dc_1742_20


Budapest University of Technology and Economics (BME)
Faculty of Electrical Engineering and Informatics (VIK)
Department of Telecommunications and Media Informatics (TMIT)
High-Speed Networks Laboratory (HSN*Lab*)
MTA-BME Information Systems Research Group

# A Function-Structure Approach to Complex Networks

D.Sc. Dissertation

In partial fulfillment of the requirements for the title of
*Doctor of the Hungarian Academy of Sciences*

András Gulyás, Ph.D.

Magyar tudósok körútja 2., H-1117 Budapest, Hungary,
E-mail: gulyas@tmit.bme.hu

Budapest
2020

*To my loving family and friends.*

# Acknowledgements

This work was carried out at the High-Speed Networks Laboratory (HSN*Lab*) at the Department of Telecommunications and Media Informatics (TMIT), Budapest University of Technology and Economics (BME) during the years 2010–2019. I am grateful to Gábor Magyar Head of the Department, for continuously supporting my research during these years.

My deepest gratitude goes to my closest collaborators, Professor József Bíró and Zalán Heszberger for the help, advice, and for those many inspiring discussions we had. My warmest thanks are due to my office mates and closest co-authors, Attila Kőrösi and Gábor Rétvári, for those hundreds of hours of talks and brainstorming we had in the last years. Grateful thanks go to the Ph.D. students I have worked with, Márton Csernai, Dávid Szabó, István Pelle and Attila Csoma. These guys have contributed in many ways (sometimes unnoticed) to the research results contained in this work. I am grateful to all colleagues of the Lab and of the Department for the friendly and inspiring atmosphere.

I wish to express my gratitude to my international collaborators Professor Dmitri Krioukov (Northeastern University, USA), Alessandra Griffa (École Polytechnique Fédérale de Lausanne (EPFL), Switzerland) and Andrea Avena-Koenigsberger (Indiana University, USA).

Exceptional thanks go to Ericsson, to MTA Bolyai Scholarship for financially supporting me and my work during my postdoc research.

Last but not least, I wish to thank my wife Gabi and my children Bandika and Nusi for their love and patience to my research-oriented lifestyle. I am grateful to my parents, Mária and László, for their care, and to my whole family, too. I wish to thank my lovely friends for all the fun we had together.

Project no. 123957, 129589 and 124171 has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the FK_17, KH_18, and K_17 funding schemes, respectively. András Gulyás was supported by the János Bolyai Fellowship of the Hungarian Academy of Sciences.

# Contents

# Chapter 1

# Introduction

If you have ever walked through a public park, you may have noticed that besides paved ways, many unpaved paths are used by people. A clear sign of this is the presence of trampled grass paths (despite the "Keep off the grass!" warnings). Modern parks are paved only after a few months of public usage, and the paving follows people's trampled paths. These paths usually unite in a visible network. People use the park in their unique way. They typically enter, and exit at various points of the park, and their behavior inside the park is also different. Some people are interested in the statues; others seek benches under shady trees or free workout areas. The network which is finally paved emerges from the summation of people's interaction with the park.

The example of public parks enlightens the nature of interaction and the co-evolution of a network and its users. In this entangled relationship, users form the structure of the web. Conversely, the emerged structure influences the behavior of the users. More abstractly, usage, or function creates structure while structure alters function. Classical studies of networks usually cover only the latter direction of the relationship. In network science [134], the observable structure of a network forms the basis of the analysis over which dynamic processes such as navigation, search, or spreading as functions hosted by the network are investigated. In this case, the structural properties of networks are modeled in a function-agnostic manner since the function is only considered after the structure is well-identified. The backward, i.e., the function $\rightarrow$ structure direction is rarely tackled in the literature. The most plausible explanation for this is that the function of a network is something tough to grasp or measure. The web of paved segments in the public park example can be easily reconstructed after a few hours or days of walking, depending on the size of the park. In fact, such maps are usually placed at the entrances showing the main attractions and roads inside the park (Figure 1.1). The map of the park acts as a kind of public information. Function, manifesting itself in the paths of people behave quite differently. The paths belong to people. The paths describe the habits of people and tell us about them. About their favorite places, the location of their homes, and even about their health (if they prefer long or short walks). The nature of the function is somewhat confidential. Some people may talk about it and give their names, others may talk about it anonymously, and others may ignore you if you ask them about their paths.

This dissertation contains models and results on the possible application and


Figure 1.1: The official map of Central Park in New York City.

benefits of the function → structure approach to complex networks. Although the behavior of the users of the network may be specific; we still can find some basic rules describing the high-level behavior of users, which gives a rudimentary characterization of function. In this work, we set out of such rules governing function and investigate networks as emergent objects coming out of the interaction of users. We show that the function → structure approach gives a complementary insight to networks compared to the widely known structure → function type studies. While the structure → function type analysis mostly reflect high-level statistics (e.g., degree distribution, clustering, diameter), the function → structure direction can identify omnipresent sub-networks or frames, and predict connection likelihood.

Although the function → structure approach may be beneficial for a broad spectrum of complex networks like biological, technological, social, or ecological networks, the rudimentary formulation of function required by the analysis is not currently available in most of the existing complex networked systems. Thus, in this dissertation, two types of networks are considered whose function can be grasped to a sufficient extent that permits the function → structure analysis. Navigable networks are a family of complex networks, over which navigation, i.e., the function of the network, can be described in terms of distributed greedy mechanisms inspired by social networks. Secondly, we investigate hierarchical networked systems in which paths, i.e., the function of the network, can be characterized by some rudimentary hierarchical relations like a customer-provider relationship. The Internet is a pathological example of such systems to which this dissertation dedicates special attention.

# Chapter 2

# An overview of complex networks

In the 18th century, the city of Königsberg, Prussia, was wealthy enough to have seven bridges across the river Pregel. The seven bridges connected four parts of lands separated by the branches of the river. The constellation is shown in Figure 2.1 where capitals (A, B, C, D) denote the lands and the bridge drawings and the corresponding handwriting (ending with the B. and Br. abbreviations) mark the location of the bridges. This scenery inspired the fantasy of the leisured inhabitants

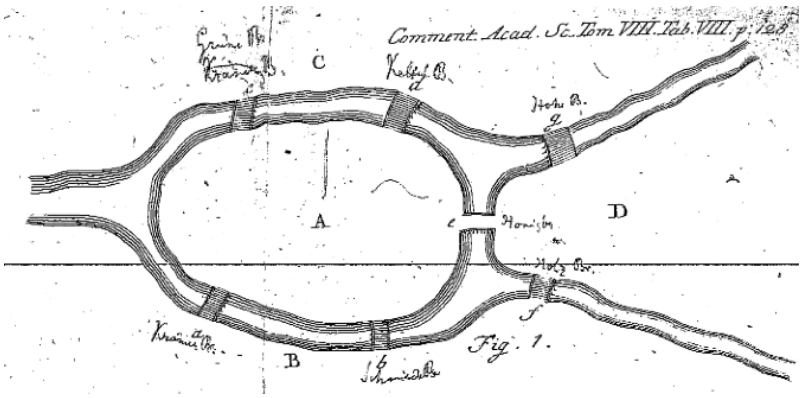



Figure 2.1: Euler's Figure 1 for the seven bridges of Königsberg problem from 'Solutio problematis ad geometriam situs pertinentis,' Eneström 53 [source: MAA Euler Archive]

of Königsberg who made a virtual playground from the bridges and lands. Their favorite game was to think about a possible walk around the bridges and lands, in which they cross over each bridge once and only once. Nobody could come up with such a fancy walk and nobody managed to prove that such a walk is impossible to find, until Leonhard Euler, the famous mathematician, took a look at the problem. Euler quickly noticed that from the perspective of the problem, most of the details of the map shown in Figure 2.1 can be omitted and a much simpler figure can be drawn focusing on the essence of the problem (see Figure 2.2).

This new representation contains only "nodes" marked with capitals (A, B, C, D) in circles representing the lands and "edges" drawn with curved lines between the nodes representing the bridges. A walk now can be described as a sequence of nodes and edges. For example the sequence A $\rightarrow$ E1 $\rightarrow$ C $\rightarrow$ E3 $\rightarrow$ D $\rightarrow$ E4 $\rightarrow$ A represents a walk starting from land A which proceeds to land C via bridge E1, then to land

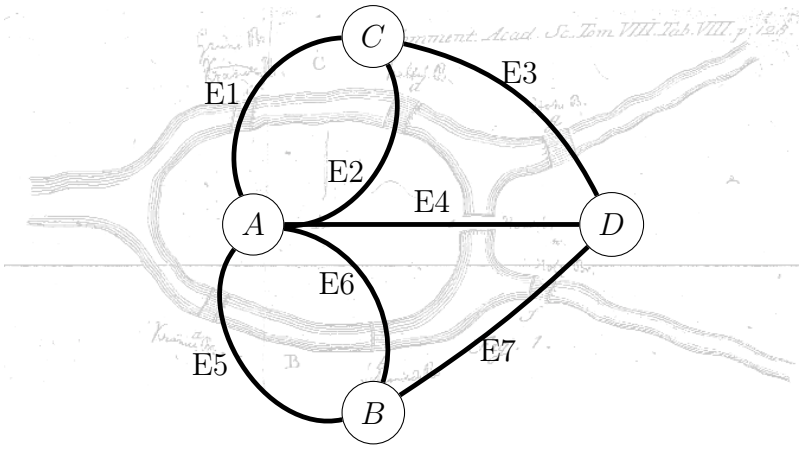



Figure 2.2: Euler's idea of abstracting away the network underlying the Seven Bridges of Königsberg puzzle.

D via bridge E3, and finally back to land A via bridge E4. All sorts of walks can be created using only the nodes and edges. All the possible walks that one can imagine throughout the bridges and lands are captured by this simple representation. The collection of nodes and edges called a network (or graph in mathematics) $G(V, E)$ turned out to be so powerful in modeling real-world problems that a whole new branch of mathematics, called graph theory has been defined based on them. In the first-ever graph-theoretic argumentation Euler showed that to find a walk crossing each bridge once and only once requires that the underlying network can contain only two nodes with an odd number of edges. In Figure 2.2, one can see that all nodes have an odd number of edges (A has five, while B, C, and D has three), which makes the problem insolvable in this network.

The network in the case of Königsberg's bridges is tiny and well-defined (contains four nodes and seven edges). Such small networks, completed with more extensive but regular networks, provided the main inputs of classical graph theory problems for around 250 years. However, the information revolution and the rapid development of digital information storage and processing technologies made it possible to gather data and analyze large, complex, and dynamic networks from all areas of life. Biological (e.g., metabolic, protein or brain networks), technological (e.g., the Internet, software, and hardware networks) and social networks (e.g., human acquaintance networks, online social networks) are the most representative examples of such complex networked systems. The need for characterization of such large and complex networks led to the definition of a wealth of network metrics, which were unknown for classical graph theory.

## 2.1 Structural properties of networks

Since the budding of network science, quite a long list of structural network properties have been defined and analyzed. However, the main resemblance of real-world networks is mostly reflected by three classes of high-level network metrics. The first class is the distance-related metrics from which diameter and average path length are of interest regarding this dissertation. The *diameter* ($D$) of a network is de-

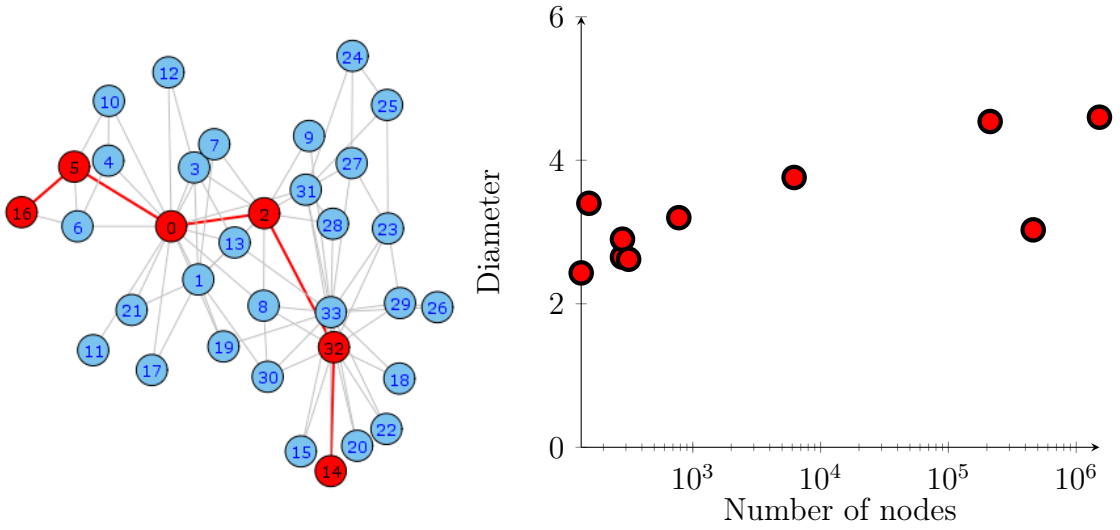



Figure 2.3: Visualization of the diameter in Zachary's karate club network [176] (left), Diameter of various real networks compared to their size (right).

fined as the length of the longest shortest path in the network. Differently put, it is the length of the shortest path between the two most distant nodes in the network (see the left panel of 2.3). Through this dissertation, we consider undirected and unweighted networks; thus, the length of a path is simply given by the number of its constituting edges. The *average path length* is the average of the lengths of shortest paths measured between all pairs of nodes in the network. Surprisingly, despite containing a very large number of nodes, the diameter and the average path length of real networks is very low. Numerous measurements [5, 57, 134] confirm that the diameter and the average path length of real networks are proportional to the logarithm of the number of nodes $N$. Such behaviour is called as the small-world property. The right panel of Figure 2.3 illustrates this relationship between network size and diameter for the Ythan estuary food web [123], Silwood park food web [123], the C. Elegans neural network [170], the E. coli, substrate graph [61], E. coli, reaction graph [61], Metabolic network of the E. coli [89], Word co-occurence network [64], MEDLINE co-authorship network [132], domain-level Internet [60] and the network of movie actors [10].

The second class of metrics captures the modular structure of the network. The most influential metric to capture network modularity is the *clustering coefficient*. Although this metric is defined in various forms in the literature [134], in this dissertation, we define the local clustering coefficient of node $i$ as :

$$c_i = \frac{\text{number of triangles connected to node i}}{\text{number of triples centered on node i}}, \tag{2.1}$$

where a "triple" means a single node with edges running to an unordered pair of others. If node $i$ has a degree of $k_i$ then the local clustering coefficient is computed in the form if:

$$c_i = \frac{2e_i}{k_i(k_i - 1)}, \tag{2.2}$$

where $e_i$ denotes the number of edges between $i$'s neighbours. The global clustering

| Network | $N$ | $\bar{k}$ | $C$ |
|---|---|---|---|
| Ythan estuary food web | 134 | 8.7 | 0.22 |
| Silwood park food web | 154 | 4.75 | 0.15 |
| C. Elegans neural network | 282 | 14 | 0.28 |
| E. coli substrate graph | 282 | 7.35 | 0.32 |
| E. coli reaction graph | 315 | 28.3 | 0.59 |
| Words co-occurence network | 460902 | 70.13 | 0.437 |
| MEDLINE co-authorship network | 1520251 | 18.1 | 0.066 |

Table 2.1: Similarities in the clustering coefficients of real networks. $\bar{k}$ denotes the average degree of the network.

coefficient is defined as the average of the local coefficients of the nodes, i.e.:

$$C = \frac{1}{N} \sum_{i=1}^{N} c_i. \tag{2.3}$$

Table 2.1 shows the striking resemblance of clustering coefficients in real networks.

Finally, the third class of metrics focuses on the variation of node degrees in the network. The degree distribution is widely used to represent the high-level structure of a system in terms of node degrees. It is defined as the distribution function $P(k)$, which gives the probability that a randomly selected node has exactly $k$ edges. Remarkably, most real networks have a power-law tail

$$P(k) \sim k^{-\gamma}, \tag{2.4}$$

where $\gamma$ is usually between 2 and 3. When it comes to visualization of the degree distribution, the complement cumulative distribution is generally used, that is $P(X > k)$. Figure 2.4 illustrates the unexpected similarity of degree distribution in networks from very diverse corners of life.

Although they are out of the scope of this dissertation, tons of other network metrics have been defined and analyzed in the literature of network science. See [48] for a nice summary of various network metrics.

## 2.2 Generative network models

Since the identification of the unexpected structural resemblance of real networks, the research community is driven by the dire need to understand the significant governing laws of network organization. One possible way of doing this is to find a set of underlying wiring mechanisms eventuating the observed high-level connectivity between the nodes of the network. Finding the appropriate wiring mechanism that generates the desired network structure casts these models as *generative models*. Most of the existing network models are qualify as generative, starting from probabilistic random graphs [29], general complex network models e.g. [11, 170], metric space models e.g. [102], fractal models [97], random walk models [29], optimization models [37] but simulation-based approaches [111, 84] are counted here too. To illustrate the philosophy of generative models, here we give a brief summary of the three most influential models of network science.

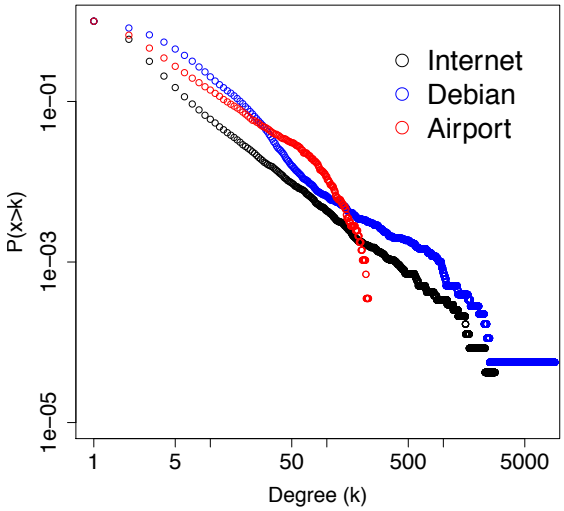



Figure 2.4: The similarity of degree distribution in three networks from diverse areas of life. The Airport network represents the network of airports and flights around the globe, the Debian network is the dependency network of packages in the Debian Linux distribution, while the Internet is the domain-level topology of the Internet.

### 2.2.1 The Erdős-Rényi (E-R) model

The most basic yet powerful network model is the random network model proposed by Pál Erdős and Alfréd Rényi. In its most simple form, all pairs of nodes are connected by a given probability $p$. Due to pure randomness, the expected number of edges in the E-R model is:

$$\bar{E} = \binom{N}{2} p = \frac{N(N-1)}{2} p, \tag{2.5}$$

thus the average degree is expected to be:

$$\bar{k} = \frac{\bar{E}}{N} = (N-1)p \approx Np, \tag{2.6}$$

for large networks, since each edge adds two degrees to the network. The E-R model thus can be easily tuned to produce realistic average degrees by setting $p = \frac{\bar{k}}{N}$.

Due to pure randomness, the small-world property of E-R graphs can be easily explained. An arbitrary node $x$ in the network will have $\bar{k}$ neighbors on average. For the neighbors of $x$ this is also approximately true. Thus, from node $x$ we will find roughly $\bar{k}^2$ nodes at a distance of two. In distance $D$ we find around $\bar{k}^D$ nodes. Thus, the diameter of the network can be implicitly given by:

$$\bar{k}^D = N, \tag{2.7}$$

from which we get $D \approx \frac{log(N)}{log(\bar{k})}$, which means that the diameter is indeed proportional to the logarithm of the number of nodes. Thus the E-R model explains the small-world property through pure randomness.

The local clustering coefficient of a given node can also be easily deduced from the properties of the model. Recall that:

$$c_i = \frac{2e_i}{k_i(k_i-1)}, \tag{2.8}$$

where $e_i$ denotes the number of edges between $i$'s neighbours. Easily, $e_i$ is expected to be:

$$e_i = \frac{k_i(k_i - 1)}{2}p, \tag{2.9}$$

thus $c_i$ is expected to be $p$, while $C \approx p$. Now this is in high contrast with the measured results seen in real networks. For an example let's see what clustering coefficient we should see if the word co-occurrence network (see Table 2.1) was completely random. In this case $\bar{k} = 70.13$ and $n = 460902$, thus we should set $p = \frac{\bar{k}}{N} = \frac{70.13}{460902} = 0.0001521582$ which means that the clustering coefficient for the corresponding random network is also 0.0001521582. In the real word network, however we see that the clustering coefficient is 0.437, which is orders of magnitude bigger (one would get similar results from any network picked from Table 2.1). This result readily proves that real networks cannot be completely random.

Finally, we show that E-R networks have a very different degree distribution from what we observe in most real-world networks. The distribution of the degree of any particular node is binomial:

$$P(k_i = k) = \binom{N-1}{k} p^k (1-p)^{1-N-k}. \tag{2.10}$$

Since

$$P(k_i = k) \to \frac{(Np)^k e^{-Np}}{k!} \tag{2.11}$$

as $N \to \infty$ and $Np =$ constant, this distribution is Poissonian. This result is also in contrast with real measurements supporting that real networks have power-law degree distribution.

### 2.2.2 The Small-world model

We have seen that random networks readily explain the small-world property exhibited by real networks, but pure randomness cannot account for their power-law degree distribution and high clustering. There is a seeming controversy between the small-world property and high clustering since making clusters arguably counteracts to small diameter. Duncan Watts and Steven Strogatz at Columbia University studied the relation of these with a very simple model [170]. Their model starts with a highly regular circular graph in which every node located in a circle layout is connected to the $K$ closest ($K/2$ in both directions) nodes on the circle. At this starting stage, the network does not exhibit the small-world property as its diameter is:

$$D \approx \frac{N}{2K}, \tag{2.12}$$

which grows linearly with $N$. However, the network has a high clustering coefficient given by:

$$C_0 = \frac{3(K-2)}{4(K-1)}, \tag{2.13}$$

independently of $N$. From the starting stage, the model iteratively goes through the nodes and take every edge connecting them to their $K/2$ rightmost neighbors and rewire it with probability $p$. Rewiring is done by replacing the endpoint of the

edge to a node picked uniformly at random from the other nodes while avoiding self-loops and link duplication. When $p = 1$, the model produces a structure close to an Erdős-Rényi random graph (see Figure 2.5).

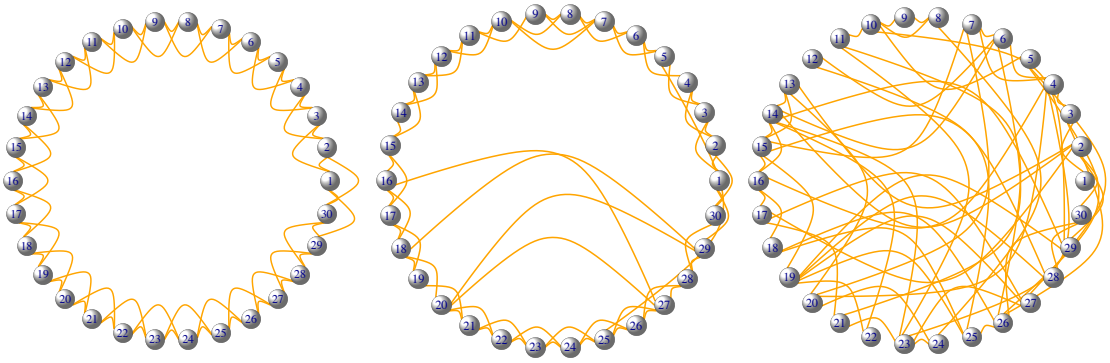


Figure 2.5: Visualization of the small-world model with $p = 0$ (left), $p = 0.05$ (middle) and $p = 0.5$ (right).

The model transforms a regular lattice having high clustering but no small-world property, into a small-world random graph with low clustering. What is interesting what happens between the two extremes, i.e., when $0 < p < 1$. Figure 2.6 shows the normalized diameter and clustering coefficient with respect to the $p = 0$ case when transiting to the completely random network at $p = 1$ in a 5000-node network. In this case, the first rewired links appear at around $p = 10^{-4}$. The diameter drops very quickly after rewiring a tiny portion of the edges. On the contrary, the clustering coefficient remains almost unaffected by these early rewirings until around $p = 10^{-2}$. There is a large space for networks exhibiting both small diameter and large clustering at $10^{-3} < p < 10^{-2}$. In this regime, the generated networks are both small-worlds and highly clusterized.

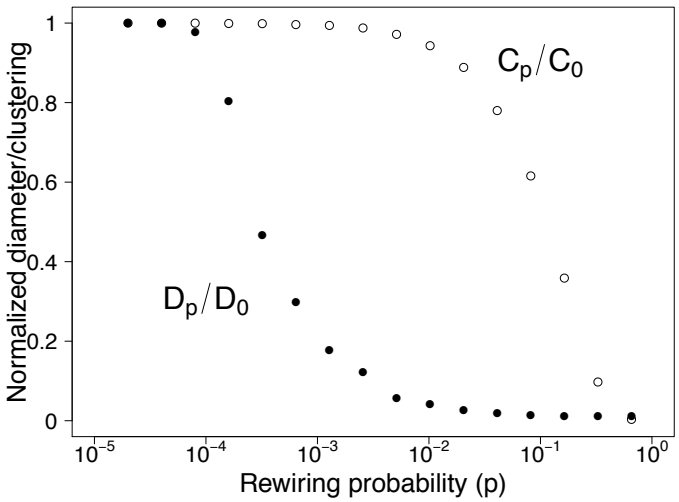



Figure 2.6: Transition from regular lattice to a random graph with the small-world model (n=5000).

Nevertheless, the degree distribution of the model is given by [16]:

$$P(k) = \sum_{\nu=0}^{\min(k-K/2,K/2)} \binom{K/2}{\nu} (1-p)^{\nu} p^{K/2-\nu} \frac{(pK/2)^{k-K/2-\nu}}{(k-K/2-\nu)!} e^{-pK/2}. \tag{2.14}$$

The shape of the degree distribution is similar to that of a random graph and has a peak at $k = K$ and decays exponentially for large $|k - K|$, which is unlike the power-law degree distribution of real networks.

### 2.2.3 The Barabási-Albert (B-A) model

We have seen that the Small-world network model can conciliate the small-world property and high clustering, however, it cannot reproduce realistic degree distribution. The most popular model in the literature capable of that was defined by Albert László Barabási and Réka Albert [11]. Their model has two main rules. First, the network is grown incrementally by adding and connecting a node at every step to the existing network, which can be an arbitrary small network in the initial phase. Secondly, the connections are not formed uniformly at random; rather, a newly arriving node tends to connect to existing nodes proportional to their degrees. More specifically, the probability that a new node connects to an existing node $i$ with a degree of $k_i$ is $p_i = \frac{k_i}{\sum_j k_j}$. This rule is called a linear preferential attachment, as the chance of the nodes present in the network to get an edge from the newly connected node grows linearly with their current degree. Figure 2.7 presents a network after 100 iterations of the model and the power-law degree distribution of a 5000-node B-A network. The average degree of the model is controlled via the fixed number of edges created by the newly arrived nodes at every step. The average path length of the B-A model is known to be growing logarithmically with the size of the network [45]. Thus B-A networks are small-worlds.

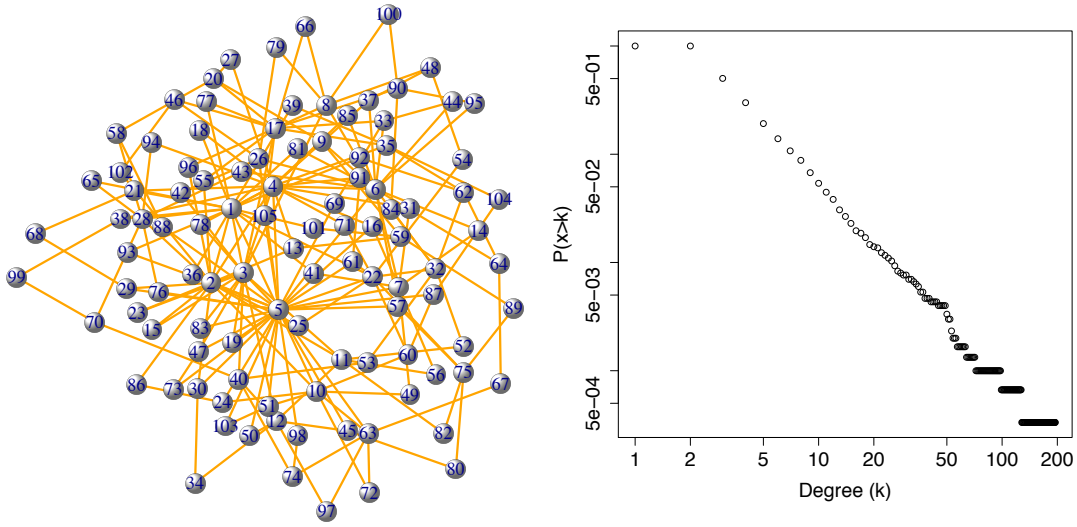



Figure 2.7: 100-node network generated with the B-A model (left) and the complement cumulative degree distribution of an 5000-node B-A network (right).

However, the clustering coefficient is proven to be inversely proportional with

the network size [96], which is unlike the high clustering coefficient independent of $n$ exhibited by real networks.

### 2.2.4 Checkpoint

Although the corresponding volume of the literature is considerable still, we can identify the following points in almost all cases of generative models:

- **Define node dynamics:** This is usually a set of rules regarding how the set of nodes residing in the network varies over time. In the simplest case, the set of nodes can be fixed a priori [29] but growth models e.g. [11] where the number of nodes increases over time seem to capture a fundamental aspect of complex networks.

- **Define edge dynamics:** This is again a rule set controlling how the edges are created between the nodes in the network. The set of rules here can be constructed in a black box fashion [29] where we want to mimic some high level edge statistics (degree distribution, diameter, clustering etc.), but also can be inspired by processes [11, 37] assumed to take place on networks.

- **Analyse and compare with measurement data:** This final and most crucial step where the outcomes of the model are computed, simulated or analytically derived and verified against the available measurement data.

While the above three-step process resulted in a great variety, often precise (as far as the power of measurement data can verify) and usable network models, the generative approach suffers basically from the inability to prove that the processes these models are defined upon, are actually there on the real network. For example, one cannot think that preferential attachment in its pure form (where a node chooses its neighbors according to their exact nodal degree) is happening in a real networked system. This inability makes the generative models and their predictions somewhat ambiguous. Such models can capture the fundamental mechanisms leading to realistic networks, but they usually cannot cope with the incentives of the nodes. Differently put, generative models readily answer the "how" question, (i.e., How real networks are generated?) but usually leave the "why" question out of scope, i.e., Why it is beneficial for nodes choosing neighbors according to a given pattern? Now, we introduce a different family of models concentrating on the incentives of the nodes and investigate the consequences of various motives.

## 2.3 Incentive-oriented models

Network formation games (NFG) constitute a nice game-theoretical framework providing considerable insight into the mechanisms that form the topology of complex real-world networks [59]. The nodes of the network are considered as selfish rational players, whose goal is to minimize costs regarding the formation of the network. More formally $\mathcal{P}$ is the set of nodes (the players in game-theory terminology) with cardinality $N$. The strategy space for node $u \in \mathcal{P}$ is to create some set of edges to other nodes in the network: $S_u = 2^{\mathcal{P}\setminus\{u\}}$. Let $s$ be a strategy vector:

$s = (s_0, s_1 \dots s_{N-1}) \in (S_0, S_1 \dots S_{N-1})$ encompassing the strategies of all nodes and $G(s)$ be the graph defined by the strategy vector $s$ as $G(s) = \bigcup_{i=0}^{N-1}(i \times s_i)$. The objective of the nodes is to minimize their cost, which is calculated as:

$$c_u = \sum_{\forall u \neq v} d^{\text{sh}}_{G(s)}(u, v) + \alpha |s_u|, \quad u, v \in \mathcal{P} \tag{2.15}$$

where $d^{\text{sh}}_{G(s)}(u, v)$ stands for the length of the shortest path between $u$ and $v$ in $G(s)$ and $\alpha$ is a constant, characterizing the cost of building an edge. With such a definition network formation games can effectively analyze the balance between link costs and distance costs as the key incentives of building specific networks structures.

In their seminal paper [59], Fabrikant et al. study the Nash equilibria (NE) of the game and show that the price of anarchy (the relative cost of the lack of coordination, PoA) can be low. In recent years there has been a flurry of research in the field of network formation games; here, we only mention a few of them. The price of stability (ratio of the best equilibrium and the social optimum, PoS) has been shown to be of $O(\log n)$ if edge costs are shared fairly among nodes [7]. Bilateral NFG and its relation to the unilateral game in the context of PoA has been studied in [46]. Some improved bounds of PoA have been presented in [4] and [52], while the latter paper has also introduced a variant of NFG, where maximum distance is used instead of the sum of distances. Michalák and Schlegel [117] improve previously known bounds on PoA in both NFG variants, and show that PoA is mostly constant in the original NFG. Network formation games also appeared in many variants since its introduction in [59]. Besides the original unweighted version, state-of-the-art literature has records for weighted games and modifications with different goals of the agents. A nice summary of the state-of-the-art can be found in [21].

While network formation games can account for the incentives of the nodes for building a specific network, it turned out that finding the appropriate incentives leading to realistic network structures is far from trivial. There have been attempts to solve this problem by altering the cost function (Eqn. 2.15) to contain structure-related parts. This flow of research enforces a particular global network structure by manipulating some terms in the local cost functions of the nodes. Several studies recovered realistic clustering and degree distribution using this technique. However, these models qualify as *exogeneous* models, where the topological constraints are explicitly built into the cost functions. Such variations of the incentive-oriented approach are very close to the philosophy of generative models. Thus the development of an incentive-oriented and *endogenous* model of network formation, that would generate more heterogeneous and realistic networks without explicitly enforcing that in the cost function, is still an open challenge [135]. We argue that the key to address this challenge is to think about how the networks are used and formalize this functioning in terms of the cost function.

# Chapter 3

# Do we pick the shortest paths in networks?

The implicit "shortest path" assumption prevailing network formation games (see Eqn. 2.15), meaning that the used communication path in a network is the one with the shortest length, also seems to dominate the network science community and most of the fundamental network metrics (diameter (Section 2.1), average path length (Section 2.1), centrality metrics [133], etc.) are computed using this assumption. Other works are supposing various models [136], network metrics (e.g. degree, centrality, congestion, homophily [156, 2, 110]) and hidden structures (e.g., hidden hierarchies and metric spaces [169, 26, 93]) guiding path selection. The contribution of these studies is remarkable in modeling and understanding path selection (alternatively routing) strategies that can recover near shortest paths without requiring global knowledge of the topology. However, a lack of confirmation with empirical data leaves an important question open: What kind of paths are *actually* chosen by nature in real-world networks?

In this chapter, we approach the question of path selection in networks from this lacking empirical angle based on [49]. Using existing and newly created datasets of the traffic flow on real-world networks, we compare the topology of the networks to the structure of empirically-determined paths extracted from these datasets. From this comparison, we infer a common characteristic of path selection in different networks called *stretch*. We present the analysis of empirically-determined paths in air transportation networks, the Internet, the fit-fat-cat word morph game, and empirically-inferred paths in the human brain. Our publicly accessible path dataset collected by the fit-fat-cat word game smartphone application is published in Scientific Data [98]. For the remainder of this chapter, we will refer to empirically-determined and inferred paths as empirical paths.

Due to their confidential nature, collecting or inferring paths in networks is a non-trivial problem. Here we list our methods capable of recording or inferring paths for every specific network in our analysis.

**Internet AS topology and real AS paths** – From the perspective of path measurements, the Internet is one of the more straightforward cases since the Internet protocol stack permits the tracing of packets using the traceroute network diagnostic tool. Although this method has its limitations [114], traceroute datasets are still the primary sources of Internet paths today. CAIDA (Center for Applied Internet

Data Analysis [163]) runs large-scale traceroute measurements regularly within the Archipelago project using the Scamper tool [113]. The recorded datasets are publicly available for download and analysis. We have downloaded a full dataset of domain-level packet traces from CAIDA, recorded on 09/29/2015, which contains around 2.5 million traces. We have also reconstructed a domain-level Internet topology based on the routing information bases of looking glass routers participating in the Route-views project [167] and the trace records of Archipelago. The obtained topology contains 52194 nodes and 117251 connections. Having both traces and the topology of the network, we were able to compare the empirical paths to their shortest possible counterparts in an approximate topology of the domain-level Internet.

**Air transportation network and flight travels** – The world's flight map is available from OpenFlights [137], from which the topology of the air transportation network can be reconstructed. For a realistic estimation of the flights used by customers, we used the Rome2Rio [148] trip planner and generated routes between 27444 randomly chosen pairs of airports. From the offered paths, we have chosen the cheapest one in the analysis. However, we note that picking according to other parameters (lowest number of transfers, lowest travel time) did not qualitatively change our results. To achieve a more realistic topology, we used airport connections extracted from Rome2Rio traces to increase the accuracy of the OpenFlight topology. The reconstructed map contained 3433 airports and 20347 flights connecting them.

**fit-fat-cat word morph game app and word chains** – For collecting paths from word networks, we have implemented a word morph game named "fit-fat-cat" for smartphones. The goal of the game is to transform a source word into a target word through meaningful intermediate words by changing only one letter at a time. The word chain fit-fat-cat is a good solution to a game with source word fit and target word cat. These word chains, collected anonymously from our users, can be considered as the footprints of human pathfinding over the word-maze of the English language. For the reconstruction of the word graph, we have downloaded the official three-letter English Scrabble words from WordFind [172] and created an edge between all the words differing only in one letter. The collected three-letter word chains were considered as our traces. For capturing only the "working" paths, we have filtered out the first 20 games (the warming up phase) and the games taking more than 30 seconds (when the players are not just using a known path but discover an unknown one) of every player. After all, we have a dataset of more than 2500 paths from 100+ players. The application is still recoding data, a current snapshot of the dataset is available from the "fit-fat-cat" public Open Science Framework data repository [99] and described in details in [98].

**Human brain and estimated paths** – Getting realistic paths from inside the human brain is tough, if not impossible. As a consequence, almost all studies in the literature concerning path-related analysis assume shortest path signaling paths. Taking into account the extreme non-triviality of path estimation in the brain, we ask here if we can use empirical anatomical and functional data to infer possible communication traces. Our dataset comprises 40 healthy human subjects who underwent an MRI session where Diffusion Spectrum Imaging (DSI) and resting-state functional MRI data were acquired for each subject. DSI data was processed following the procedures described in [80, 34, 51], resulting in 40 weighted, undirected structural connectivity maps ($GS$) comprising 1015 nodes, where each node repre-

sents a parcel of cortical or subcortical gray matter, and connections represent white matter streamlines connecting a pair of brain regions. Connection weights determine the average density of white matter streamlines and here only consider connections with density above 0.0001, resulting in $GS$ with an average of 12596.2 connections per subject. Functional MRI data were processed following state of the art pipelines described in [127, 145], yielding a BOLD signal time-series per node, each with 276 points that were sampled every 1920 ms. The magnitude of the BOLD signal is an indicator of the degree of neural activity at a node. Combining structural and functional data, we infer feasible structural pathways through which neural signals might propagate using the following process. $(i)$ Identify source-destination pairs with high statistically-dependent brain activity. We searched for pairs of nodes such that the Pearson correlation of the BOLD signal time series - without global regression - was above 90%. These nodes were used as the source-destination pairs of our paths. $(ii)$ Determine which nodes are active at every time-step. We say that a node is "active" at a given time-step if the BOLD signal is $> \gamma$ and "inactive" otherwise. We construct activity vectors for each time-step indicating which nodes were active. Here we use $\gamma = 0$, but we get qualitatively similar results for near-zero $\gamma$. $(iii)$ Construct subgraphs of active nodes. We constructed a subgraph $GS_i$ of $GS$ for each time-step by considering only the nodes that are active at a given time-step $i$. $(iv)$ Define paths between source-destination node pairs. For all of our source-destination pairs (generated in step $(i)$), we considered the shortest path in the $GS_i$ graphs, if the path existed. If there were multiple shortest paths between a source-destination pair, we choose one randomly. Our source-destination traces include the paths found across all $GS_i$ subgraphs. It is worth noting that this method assumes that information can only traverse active nodes. Furthermore, we are considering here a model for large spatial and temporal scale communication in brain networks that is not necessarily applicable to neural networks at smaller scales. While we cannot validate with empirical data whether these paths are actually used for the flow of neural signals, from a path inflation perspective, we can consider these paths as a lower bound on the length of the real signaling pathways.

The main topological features of our networks and the statistics of the empirical paths are shown in Table 3.1.

| Network | Airport | Intern. | Brain | fit-fat-cat |
|---|---|---|---|---|
| # Nodes | 3433 | 52194 | 1015 | 1015 |
| # Edges | 20347 | 117251 | 12596.2 | 8320 |
| Avg. deg. | 11.85 | 4.49 | 24.82 | 16.39 |
| Avg. clust. | 0.64 | 0.32 | 0.42 | 0.44 |
| Avg. dist. | 3.98 | 3.93 | 2.997 | 3.52 |
| Diam. | 13 | 11 | 6.4 | 9 |
| # Emp. paths | 13722 | 2422001 | 394072 | 2700 |
| Path avg. dist. | 4.67 | 4.21 | 4.16 | 3.82 |

Table 3.1: Basic structural properties of our networks and paths we have analyzed.

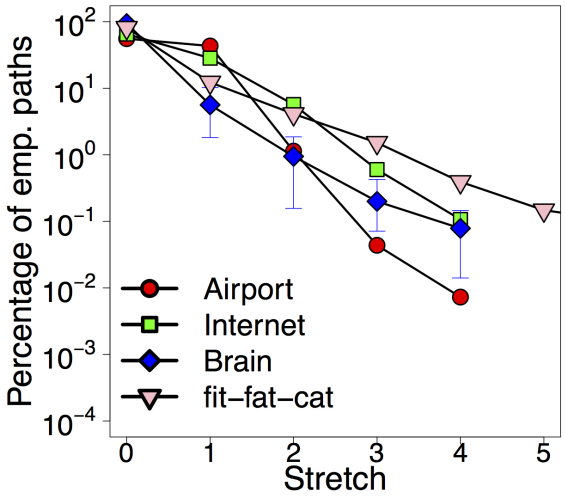



Figure 3.1: Stretch of the empirical paths with respect to their shortest counterparts. While most of the empirical paths exhibit zero stretch (confirming the shortest path assumption), a large fraction (20-40%) of the paths is "inflated" even up to 4-5 hops. The plot indicates a stunning resemblance in the distribution of path stretch in our networks.

Our striking finding based on the datasets is that traffic in networks does not necessarily follow the shortest paths. Fig. 6.9 presents the stretch of the paths, which is computed as the length of the empirical path minus the length of the corresponding shortest path having the same source and destination pair. The figure shows a significant resemblance in the distribution of path stretch across our networks. While around 60-80% of the empirical paths exhibit zero stretches, the remaining paths show path stretch which can exceed up to 4-5 hops. From this result, two things follow. First, the plot confirms the shortest path assumption of previous studies in the sense that most of the empirical paths are shortest indeed. In this respect, nature's path selection policy definitely "prefers short paths". However, the non-negligible portion (20-40%) of inflated paths suggests that there may be other navigational policies at use simultaneously. The main topological features of our networks and the statistics of the empirical paths are shown in Table 3.1.

## 3.1 Navigability: The primary function of networks

The most plausible explanation of the above-experienced stretch is that real pathfinding algorithms somewhat differ from the shortest path algorithm. While an algorithm in possession of the global topology of the network can easily find the shortest path between arbitrary pairs of nodes, real entities having localized views of the network need to operate differently when navigating in the wild between nodes (see Figure 3.2). Milgram's famous experiment [165] was the first empirical evidence that complex networks are navigable by distributed greedy search algorithms. Besides the remarkable approximation of the diameter of the social network, Milgram showed that people can effectively navigate through their social acquaintances without knowing the structure of the complete network of social interactions. This

experiment triggered the extensive study of the navigational aspect of complex networks.

Networks are efficient conduits of information and other media. News, ideas, opinions, rumors, and diseases spread through social networks fast, sometimes becoming viral for reasons that are often difficult to predict [15, 91, 169, 142, 147, 62, 56, 116, 17, 121, 67, 125]. Many biological networks are also paradigmatic examples of information routing, ranging from information processing and transmission in the brain, to signaling in gene regulatory networks, metabolic networks, or protein interactions [13, 174, 30, 40]. Perhaps the most basic example is the Internet whose primary function is to route information between computers. If one is to list some common functions of different networks, then information routing will likely be close to the top. It is thus not surprising that many networks were found navigable, meaning that nodes can efficiently route information through the network even though its global structure is not known to any individual node [118, 166, 92, 53, 109, 156, 26, 36, 85, 105, 106, 35].

Recent research efforts suggest that presuming the existence of a hidden metric space behind complex networks seems a plausible explanation for their excellent navigational properties [26, 102]. A distributed greedy navigation algorithm always makes the locally optimal choice at each step, hoping that this will result in a global optimum or its sufficiently good approximation. If greedy navigation can lean on a metric space during the search, this hope becomes a reality with high probability (see Figure 3.2). For example, in his pioneering model [93] Kleinberg effectively used the $D$-dimensional euclidean lattice to describe navigation in small worlds, while for the explanation of Milgram's experiment, a hierarchical model has been proposed in [169]. Since metric spaces are either existent [169] or can be efficiently constructed concerning social and computer networks [25, 95], greedy navigation is a remarkably efficient strategy in finding network paths. Furthermore, many practical routing solutions are based on the greedy navigation principle. Perhaps the most successful practical systems using greedy forwarding are the overlay networking solutions based on distributed hash tables, e.g., CAN [152] and Chord [159]. These schemes employ different underlying abstract geometries, torus [152] and circle [159] respectively, as a basis of forwarding. In [160] several greedy navigation schemes for ad hoc networks, based on geographical distance, are surveyed. Hamming-distance based greedy navigation has been utilized in Microsoft's BCube data center design [32]. In the field of cognitive neuroscience, recent studies reported major correlations between navigation and learning skills of humans [128, 122] while others go even further and investigate the possibility that navigation in cognitive spaces may lie in the core of any form of organized knowledge and thinking [20, 58, 19].

Our observations with empirical paths and the wealth of studies in the literature suggest that instead of shooting for the shortest possible path, real path selection mechanisms use greedy navigation supported by a physical or cognitive metric space. Since the primary purpose of a network is to enable traffic exchange between its constituting nodes somehow, navigability seems to be the primary function real networks host in the first place. Now we will start from the functionality of navigation and try to deduce the structure of the network as a consequence of hosting navigability. Thus in the following, the function $\rightarrow$ structure approach could be either translated into navigation $\rightarrow$ structure, but we reserve the framework to be

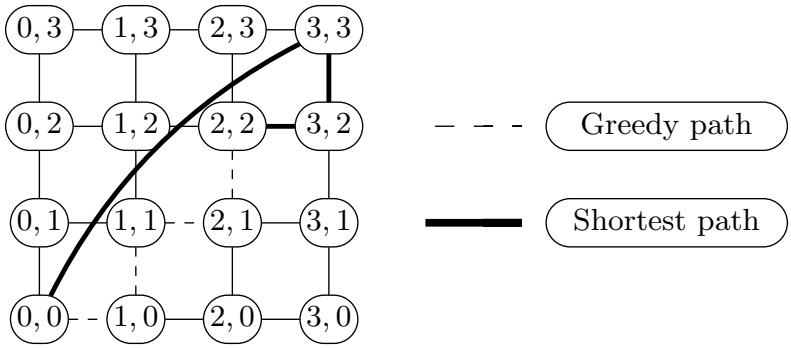



Figure 3.2: Deviation of shortest and greedy paths in the 2D Euclidean grid between nodes $(2, 2)$ and $(0, 0)$. In possession of the complete topology of the network, the shortest path algorithm can easily find the shortest (represented by a continuous line) path. Greedy navigation relies on the metric properties of the grid. When starting at node $(2, 2)$, the greedy navigation algorithm computes the distances of the possible next hops (i.e. nodes $(1, 2), (2, 3), (3, 2), (2, 1)$) from the destination $(0, 0)$. Nodes $(1, 2)$ and $(2, 1)$ are found to be 3, while nodes $(2, 3)$ and $(3, 2)$ to be 5 steps away. Thus the greedy navigation algorithm picks a local optimum and chooses randomly between the two closest nodes (i.e. $(1, 2)$ and $(2, 1)$, in the illustrated case its node $(2, 1)$. The global optimum in this case would be node $(3, 2)$. After stepping to node $(2, 1)$, the greedy navigation algorithm draws a similar decision based on distance computations in the 2D grid.

able to handle functions other than greedy navigability (see an argument for hierarchical systems in Chapter 6). More specifically, we will set up and analyze an incentive-oriented model, capable of handling the function of navigability in its cost functions.

# Chapter 4

# From function to structure in navigable networks

The finding that real networks are navigable with greedy algorithms does not necessarily mean that they evolve to become navigable. Navigability can be a by-product of some other evolutionary incentives because different networks have many other different functions as well. In other words, it remains unclear if ideal networks whose only purpose is to be maximally navigable at minimal costs have anything in common with real networks. Even if they do, then how close are real networks to these ideal maximally navigable configurations? If they are close but not exactly there, or if their navigability suddenly deteriorates, possibly signifying the onset of a disease [47], then what can we do to cure the network and boost its navigability?

Here we show that the ideal maximally navigable networks do share some basic structural properties with the Internet, *E.coli* metabolic network, English word network, US airport network, the Hungarian road network, and a structural network of the human brain. Yet these ideal networks are not generative models of the real networks, where by generative models we mean function-agnostic models that simply try to reproduce some structural properties of real networks. Instead, these ideal networks coming out of function-based incentive-oriented models identify minimal sets of edges that are most critical for navigation in the real network. In other words, they are navigation skeletons or subgraphs of real networks. We find that the considered real networks contain high percentages, exceeding 90% in certain cases, of edges from their navigation skeletons, while the probability of such containment in randomized null models is exponentially small. The knowledge of these skeletons allows us to quantify precisely what connections the considered real networks lack to be maximally navigable, and which of their connections are not exactly necessary for that. To define and construct these maximally navigable network skeletons, we employ game theory.

Game theory is a standard tool to study the behavior of a population with given incentives. The population members are called players, and their possible actions are strategies, while cost functions or payoffs express players' incentives. The purpose of a player is to minimize her costs (or maximize her payoffs) by adjusting her strategy. A Nash equilibrium is a game state such that no player can further reduce her costs by altering her strategy unilaterally. Such equilibrium states are local optima where the game can eventually settle after some transient dynamics. The global optimum

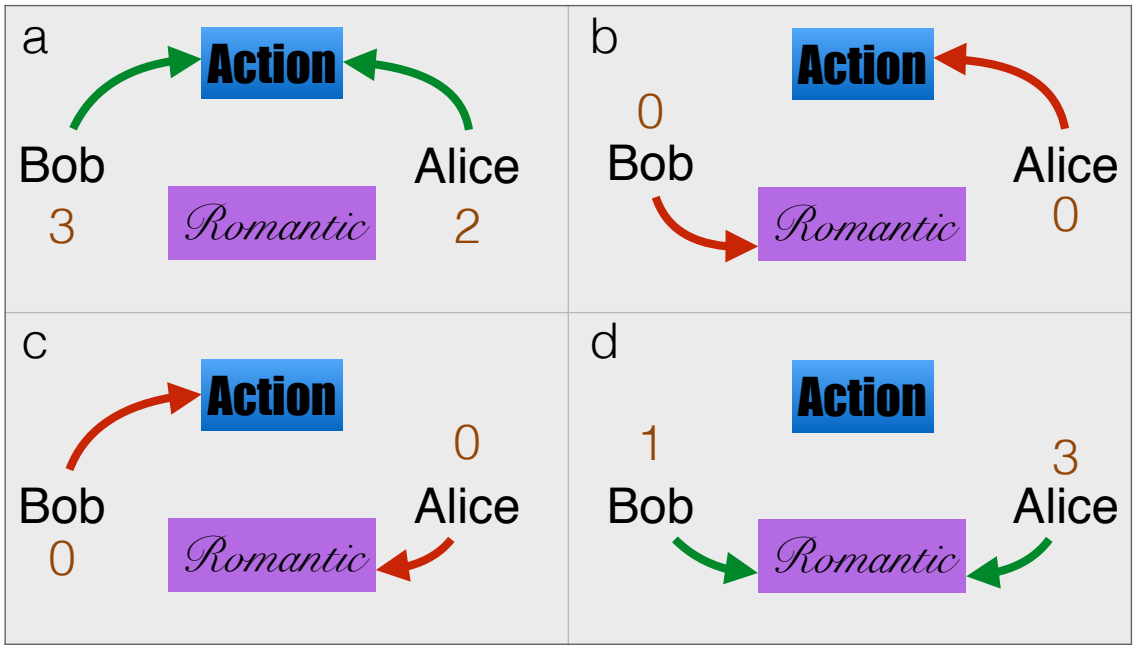



Figure 4.1: Illustration of game theory. Alice and Bob are happy only if they go out to the movies together, but the level of their happiness depends on what movie they watch. The basic notions of a game: Players: Alice and Bob; Strategies: Go to see an action or a romantic movie; Payoffs: The level of happiness $0, 1, 2, 3$; Nash equilibria: situations in which the players cannot be happier by unilaterally modifying their strategies. In the figure, states (a) and (d) are equilibria when Alice and Bob go together to watch a movie. State (a) is the global optimum since the total happiness $3 + 2 = 5$ is maximized.

is an optimum where the total cost of all players is minimized. Since the inception of game theory, a broad palette of games has been introduced, modeling diverse properties of real-life situations [135], Figure 4.1.

Here we use game theory to find the structure of networks that are Nash equilibria of a network construction game [135, 59, 7, 46, 4, 52, 117] with navigability incentives. The concept of Nash equilibrium captures the idea of self-organization, i.e., of the emergence of structures from the local interaction of rational but selfish players, in contrast with global optimization used in centralized planning of globally optimal navigable structures [104]. In our Network Navigation Game (NNG), players are network nodes whose optimal strategy is to set up a minimal number of edges to other nodes ensuring maximum navigability. That is, the cost function reflects trade-offs between the number of created edges and navigability. If each node connects to every other node, then this construction is maximally navigable but maximally expensive, too. If no edges are set up, then the cost is zero, but so is navigability. There is a sweet spot of the least expensive but still 100%-navigable network, defined as the network in which all pairs of nodes can successfully communicate using geometric routing [138]. The goal of our game is to find this sweet spot.

The network construction game that we employ is very general and applies to any set of points in any geometry. The latent geometry of numerous real networks is not

Euclidean but hyperbolic, as shown in [141]. Specifically, the model in [141] extends the preferential attachment mechanism of network growth by observing that in many real networks, the probability of establishing a connection depends not only on the popularity of nodes, i.e., their degrees but also on the similarity between nodes. The similarity is modeled in [141] as a distance between nodes on the simplest compact space, the circle. The connection probability thus depends both on node degrees (popularity) and on the distance between nodes on the circle (similarity). The node degrees are then mapped to radial coordinates of nodes, thus moving nodes from the circle to its interior, the disk. One can then show that the resulting connection probability depends only on the hyperbolic (versus Euclidean) distance between nodes on the disk, and that the resulting graphs are random geometric graphs [144] growing over the hyperbolic plane. As shown in earlier work [101], these graphs are maximally random, i.e., maximum-entropy graphs that have power-law degree distributions and strong clustering. In other words, power-law degree distributions, coupled with strong clustering, are manifestations of latent hyperbolic geometry in networks. If this geometry is not hyperbolic but Euclidean, then the resulting random geometric graphs still have strong clustering, but their degree distributions are Poisson distributions that do not have any fat tails [144]. The model in [141] has been validated against long histories of growth of several real networks, predicting their growth dynamics with remarkable precision. It is then not surprising that, as a consequence, the same model also reproduces a long list of structural properties of these networks [141].

Random geometric graphs [144] are defined as sets of points sprinkled uniformly at random over a (chunk of) geometric space. Every pair of points is then connected if the distance between the points in the space is below a certain threshold. Given that the latent space of real scale-free networks is hyperbolic, our starting point is the first part (uniform sprinkling) of the random geometric graph definition. That is, we first randomly sprinkle a set of points over a hyperbolic disk. We then do not proceed to the second part of the random geometric graph definition. Instead, given only the coordinates of sprinkled nodes, we identify the sets of edges, ideal for navigation, that corresponds to the Nash equilibria of our NNGs. We then analyze the structural properties of the resulting ideal-navigation networks and find that, surprisingly, they also have power-law degree distributions and strong clustering. This result invites us to investigate if these navigation-critical edges exist in real networks. To check that, we have to know the hyperbolic coordinates of nodes in these real networks in the first place. We infer these coordinates in the considered collection of real networks using the deterministic HyperMap algorithm (Section 5.6). Given only these inferred coordinates, we then construct the ideal-navigation Nash equilibria defined by these coordinates, and compare, edge by edge, the resulting Nash equilibrium networks against the real networks. We find that the real networks contain large percentages of edges from their Nash equilibria. This methodology thus allows us to identify the navigation skeleton of a given real network. We finally check directly that edges in these skeletons are indeed most critical for navigation by showing that their alterations affect drastically network navigability.

## 4.1 Definition of the function-structure approach for navigable networks

We start with a set of nodes $u = 1, 2, \ldots, N$, i.e., $N$ nodes, scattered randomly over a hyperbolic disk of radius $R$. The densities of nodes' polar coordinates $(r, \phi)$, $r \in [0, R]$, $\phi \in [0, 2\pi]$, are [101]

$$\rho(r) = \frac{\alpha \sinh(\alpha r)}{\cosh(\alpha R) - 1}, \quad \rho(\phi) = \frac{1}{2\pi}, \tag{4.1}$$

where $\alpha > 1/2$ is a parameter controlling the heterogeneity of the layout. If $\alpha = 1$, the nodes are distributed uniformly over the hyperbolic disk because the area element at coordinates $(r, \phi)$ is $dA = \sinh(r)\, dr\, d\phi$. The desired node scattering is achieved in simulations by placing nodes $u$ at polar coordinates $r_u = (1/\alpha)\,\mathrm{acosh}\,\{1 + [\cosh(\alpha R) - 1]\, U\}$ and $\phi_u = 2\pi U$ where $U$ for each $u$ is a random number drawn from the uniform distribution on $[0, 1]$. The hyperbolic distance between any two nodes $u$ and $v$ is

$$d(u, v) = \mathrm{acosh}\left[\cosh r_u \cosh r_v - \sinh r_u \sinh r_v \cos(\phi_u - \phi_v)\right]. \tag{4.2}$$

In greedy geometric routing, node $u$ routes information to some remote node $v$ by forwarding the information to its connected neighbour $u'$ closest to $v$ in the plane according to the distance above. If $u$ has no neighbour $u'$ closer to $v$ than $u$self, then navigation fails, and we say that $u$ cannot navigate to $v$. The percentage of pairs of nodes $u, v$ such that $u$ can successfully navigate to $v$ is called the success ratio. If this percentage is 100%, we say that the network is maximally (100%) navigable.

The strategy space of node $u$ is all possible combinations of edges that $u$ can establish to other nodes. One extremal strategy is to establish no edges. The other extreme is to connect to everyone. The total number of possibilities for $u$ is $2^{N-1}$. Any combination of strategies that all nodes select is a network on $N$ nodes.

The objective of each node $u$ is to set up a minimal number of edges to other nodes such that $u$ can still navigate to any other node in the network. Formally, the cost function of node $u$ that it minimises is $c_u = k_u + n_u$, where $k_u$ is the number of edges that $u$ establishes, and $n_u$ is either zero if $u$ can navigate to everyone, or infinity otherwise. A more formal description of the strategies and payoffs can be found in Figure 4.2.

Given any node $u$, we call node $v$'s coverage area the set of all points closer to $v$ than to $u$ (see Figure 4.3). Trivially $v$ covers itself, since it is closer to itself ($d(v, v) = 0$) than to $u$. Therefore if $u$ connects to all other nodes, then $u$ trivially covers them all. The optimal strategy for $u$ minimizing $u$'s costs is thus to connect to a minimal number of nodes such that the union of their coverage areas contains all the other nodes. Indeed, if $u$ does that, and if all other nodes do the same, then the resulting network is 100%-navigable at the minimal number of edges. The network is fully navigable because if $u$ wants to navigate to any remote node $w$, then by construction, there exists $u$'s neighbour $v$ that contains $w$ in its coverage area, and $u$ can use $v$ as the next hop towards $w$. If $v$ is not directly connected to $w$, then there exists $v$'s neighbour $v'$ that contains $w$ in its coverage area, so that $v$ can route to $v'$, and so on until the information reaches destination $w$ lying within the intersection

**Strategies.** The strategy space for a node $u \in \mathcal{P}$ is to create some set of arcs to other nodes in the network: $S_u = 2^{\mathcal{P}\setminus\{u\}}$. Let $s$ be a strategy vector: $s = (s_0, s_1 \dots s_{N-1}) \in (S_0, S_1 \dots S_{N-1})$ and $G(s)$ be the graph defined by the strategy vector $s$ as $G(s) = \bigcup_{i=0}^{N-1}(i \times s_i)$.
**Payoff.** The objective of the nodes is to minimise their cost which is calculated as:

$$c_u = \sum_{\forall u \neq v} d_{G(s)}(u, v) + |s_u|, \quad u, v \in \mathcal{P} \tag{4.3}$$

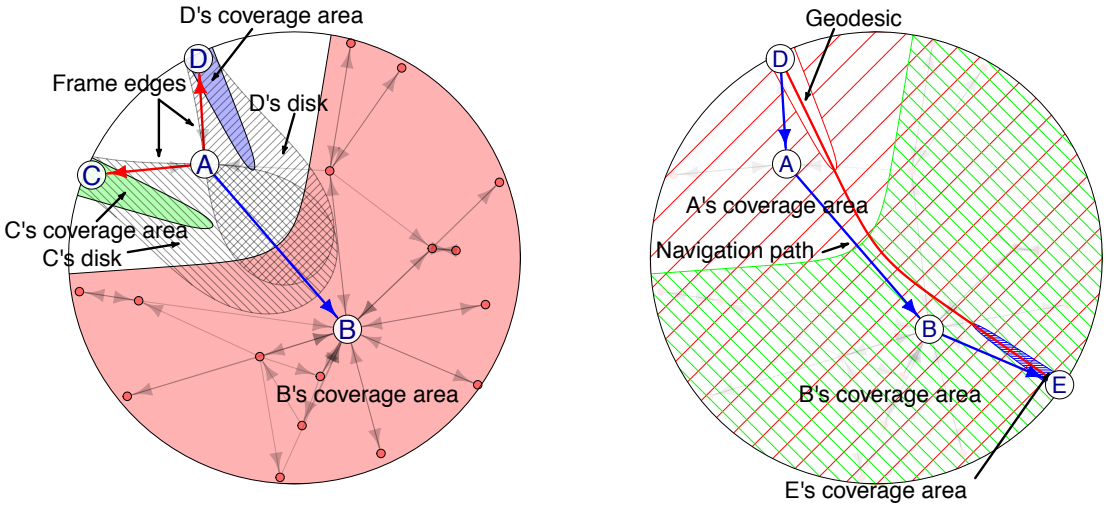
where

$$d_{G(s)}(u, v) = \begin{cases} 0 & \exists\ u \to v \text{ greedy path in } G(s) \\ \infty & \text{otherwise.} \end{cases}$$

Figure 4.2: Formal definition of the Network Navigation Game (NNG)

of all the coverage areas along the path (Figure 4.3). The problem of finding the optimal set of edges for $u$ thus reduces to the minimum set cover problem [70]. More formally The Nash equilibria of the Network Navigation Game can be characterized for each node independently as follows: take a node $u$, and for all $v \in V \setminus u$ let $\mathbb{S}_v^u = \{w | d(v, w) < d(u, w)\}$. Trivially $\mathbb{S}_v^u \subset V$ and $\bigcup_{v \in V\setminus u} \mathbb{S}_v^u = V$. The optimal strategy $s_u^{\text{opt}}$ of $u$ is the minimal set cover of $V$ with the sets $\mathbb{S}_v^u$, independently from the strategies of the other nodes. This means that $s = (s_1^{\text{opt}}, s_2^{\text{opt}} \dots s_{N-1}^{\text{opt}})$ is both a NE and a social optimum.

The Nash equilibrium is not necessarily unique as there can exist different solutions of the above set cover problem. In our work, we concentrate on a specific equilibrium, which, besides being a solution, it also minimizes the sum of edge the edge lengths all over the network. This is fully in line with the edge-locality principle of complex networks [169] [93] [26] which many times accounted for the high clustering coefficient. More formally, from the strategy vectors constituting a Nash equilibria $s_i$ and the corresponding graphs $G(s_i) = \bigcup_{i=0}^{N-1}(i \times s_i) = (V, E_i)$ we seek for the one minimising $\sum_{j \in E_i} d(E_i(j))$.




(a) (b)

Figure 4.3: Illustration of the network navigation game (NNG). Panel (a) shows the optimal set of connections (optimal strategy) of node A in a small simulated network. All nodes are distributed uniformly at random over the hyperbolic disk, and A's optimal strategy is to connect to the smallest number of nodes ensuring maximum (100%) navigability. These nodes are B, C, and D because it is the smallest set of nodes whose coverage areas, shown by the colored shapes, contain all other nodes in the network. B's coverage area for A (red) is defined as a set of points hyperbolically closer to B than to A, therefore if A is to navigate to any point in this area, A can select B as the next hop, and the message will eventually reach its destination, as the second panel illustrates. Link AC (and AD) in panel (a) is also a frame link, because A is the closest node to C, as illustrated by the hyperbolic disk of radius $|AC|$ centred at C (the line-filled shape), which does not contain any nodes other than C and A. Therefore to navigate to C, A has no choice other than to connect directly to C. Panel (b) shows the sequence of shrinking coverage areas along the navigation path (blue arrows) from D to E. The red curve is the geodesic between D and E in the hyperbolic plane. The coverage areas are shown by the shapes filled with lines of increasing density. The largest is A's coverage for D. The next one is B's coverage for A. The smallest is E's coverage for B.

## 4.2 An Euclidean example

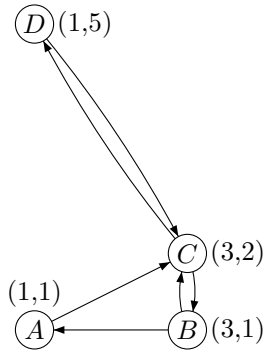



Figure 4.4: Network Navigation Game in the Euclidean plane.

As an example, let us compute the Nash equilibrium topology for four points in the Euclidean plane $A, B, C, D$ (see Figure 4.4). Any node $u$ out of these four needs to have a greedy next hop towards any other nodes (to avoid infinite cost) while having its number of edges minimized. Note that having a greedy next hop is sufficient since all the other nodes will have greedy next hops towards any other nodes for ensuring $c_u \leq \infty$, which implies greedy paths between arbitrary pairs of nodes.

Let us compute the sets $\mathbb{S}_v^u = \{w | d(v,w) < d(u,w)\}$ for the nodes, where $d(x,y)$ is the Euclidean distance and the minimal set covers for each node to get the Nash equilibrium.

- $\mathbb{S}_A^B = \{w | d(A,w) < d(B,w)\} = \{A, D\}$, which means that $A$ is a good greedy next hop towards $A$ and $D$ for $B$, similarly $\mathbb{S}_C^B = \{C, D\}$, $\mathbb{S}_D^B = \{D\}$ therefore the minimal cover for $B$ is {A,C} so $B$ creates two edges to $A$ and $C$
- $\mathbb{S}_B^A = \{B, C\}$, $\mathbb{S}_C^A = \{C, B, D\}$, $\mathbb{S}_D^A = \{D\}$ therefore the minimal cover for $A$ is {C} so $A$ creates one edge to $C$
- $\mathbb{S}_A^C = \{A\}$, $\mathbb{S}_B^C = \{B, A\}$, $\mathbb{S}_D^C = \{D\}$ therefore the minimal cover for $C$ is {B,D} so $C$ creates two edges th $B$ and $D$
- $\mathbb{S}_A^D = \{A, B, C\}$, $\mathbb{S}_B^D = \{B, C, A\}$, $\mathbb{S}_C^D = \{C, B, A\}$ therefore the minimal cover for $D$ is for example {C} ($A$ and $B$ would be also good) so $D$ creates one edge to $C$

Thus we can construct the graph from these minimal set coverings see Figure 4.4. This is a Nash equilibrium and a social optimum as there are no lower cost equilibria or state for this game.

## 4.3 Reformulation of the problem using statistical mechanics

The network navigation game for navigable networks above grasps many aspects of the function $\rightarrow$ structure approach to networks. The navigational incentive of the

nodes manifests the function of the network while the game-theoretical formulation incorporates self-organization. A major problem, however, with this formulation is raised by computability. The set-cover problem is NP-hard [90] and the network construction game relies on solving $N$ instances of that. Thus, the solution of the construction game in its current form prohibits theoretical analysis and tractable numerically only up to a few thousand nodes with contemporary software solvers. To grasp high-level statistics of the equilibrium networks, we recast the problem using statistical mechanics. Although we loose capturing the exact solution to a given constellation, what we gain is a high-level analysis valid even in the limit $\lim_{N\to\inf}$. We prove the validity of the analysis via extensive computer simulations.

### 4.3.1 A brief overview of the statistical mechanics of networks

Statistical mechanics is proven to be a useful theoretical toolset for the analysis of complex networks [43]. Models of networks based on statistical mechanics are ensemble models, meaning that a model is defined to be not a single network, but a probability distribution over many possible networks. Using this approach, the goal of our analysis will be to choose a probability distribution such that networks that are a better fit to observed characteristics (e.g., the consequences of navigability) are given higher probability in the model. Consider the set of all simple graphs without self-loops and multiple edges on $N$ vertices $\mathcal{G}$ of graphs.

Suppose we have a collection of graph observables $x_i, i = 1...r$, that we have measured in empirical observation of some real-world network or networks. We also assume that we have an estimate $\langle x_i \rangle$ of the expectation value of each observable. Let $G \in \mathcal{G}$ be a graph in our set of graphs and let $P(G)$ be the probability of that graph within our ensemble. We want to choose $P(G)$ so that the expectation value of each of our graph observables $x_i$ within that distribution is equal to its observed value, but in any other aspect, it ensures maximal randomness. This condition is fulfilled if maximizing the Gibbs entropy,

$$S = -\sum_{G\in\mathcal{G}} P(G) \ln P(G), \tag{4.4}$$

subject to the constraints

$$\sum_{G} P(G) x_i(G) = \langle x_i \rangle, \tag{4.5}$$

and the normalization condition

$$\sum_{G} P(G) = 1. \tag{4.6}$$

Here $x_i(G)$ stands for the value of $x_i$ in graph $G$. By assigning Lagrange multipliers $\alpha, \theta_i$, we then find that the maximum entropy is achieved for the distribution satisfying

$$\frac{\partial}{\partial P(G)} \left[ S + \alpha \left( 1 - \sum_{G} P(G) \right) + \sum_{i} \theta_i \left( \langle x_i \rangle - \sum_{G} P(G) x_i(G) \right) \right] = 0 \tag{4.7}$$

for all graphs $G$. This gives

$$\ln P(G) + 1 + \alpha + \sum_i \theta_i x_i(G) = 0, \tag{4.8}$$

or equivalently

$$P(G) = \frac{e^{-H(G)}}{Z}, \tag{4.9}$$

where $H(G)$ is the graph Hamiltonian

$$H(G) = \sum_i \theta_i x_i(G), \tag{4.10}$$

and $Z$ is the partition function

$$Z = e^{\alpha+1} = \sum_G e^{-H(G)}. \tag{4.11}$$

It is useful to define the free energy

$$F = -\ln Z, \tag{4.12}$$

derivatives of which specify a system of equations between the Lagrange multipliers and the observables

$$\frac{\partial F}{\partial \theta_i} = \langle x_i \rangle. \tag{4.13}$$

Equations 4.9 4.10 and 4.9 define the so called exponential random graph model. The expected value of any graph property $x$ within the model is simply

$$\langle x \rangle = \sum_G P(G)x(G). \tag{4.14}$$

For an example, consider one of the simplest of exponential random graphs having $N$ nodes. Our observation is only the expected number of edges $\langle m \rangle$ that our network should have. In that case the Hamiltonian takes the simple form

$$H(G) = \theta m(G). \tag{4.15}$$

We define the adjacency matrix $\boldsymbol{\sigma}$ to be the symmetric $N \times N$ matrix with elements

$$\sigma_{ij} = \begin{cases} 1, & \text{if } i \text{ is connected to } j \\ 0, & \text{otherwise.} \end{cases} \tag{4.16}$$

Then the number of edges is $m = \sum_{i<j} \sigma_{ij}$, and the partition function is

$$Z = \sum_G e^{-H(G)} = \sum_{\sigma_{ij}} \exp\left(-\theta \sum_{i<j} \sigma_{ij}\right) = \prod_{i<j} \sum_{\sigma_{ij}=0}^{1} e^{-\theta\sigma ij} = \prod_{i<j} (1+e^{-\theta}) = \left[1 + e^{-\theta}\right]^{\binom{N}{2}}. \tag{4.17}$$

The free energy in this case is

$$F = -\binom{N}{2} \ln(1 + e^{-\theta}). \tag{4.18}$$

Then, for instance, the expected number of edges in the model is simply

$$\langle m \rangle = \frac{1}{Z}\sum_{G} m e^{-H} = -\frac{1}{Z}\frac{\partial Z}{\partial \theta} = \frac{\partial F}{\partial \theta} = \binom{N}{2}\frac{1}{e^{\theta}+1}. \tag{4.19}$$

We can also express the parameter $\theta$ in terms of

$$p = \frac{1}{e^{\theta}+1}, \tag{4.20}$$

so that $\langle m \rangle = \binom{N}{2}p$. The probability $P(G)$ of a graph in this ensemble can be written as

$$P(G) = \frac{e^{-H}}{Z} = \frac{e^{-\theta m}}{[1+e^{-\theta}]^{\binom{n}{2}}} = p^{m}(1-p)^{\binom{N}{2}-m}. \tag{4.21}$$

In other words, $P(G)$ is the probability for a graph in which each of the $\binom{n}{2}$ possible edges appears with an independent probability $p$. This model is basically identical to the Erdős-Rényi model defined in section 2.2.1.

Instead of specifying a Hamiltonian for a global quantity like the expected number of edges, we can define one coupling to each edge. In this case, we suppose that the observables are the edges, i.e., the elements of the adjacency matrix $\sigma_{ij}$ and the model constraints are defined on $\langle a_{ij} \rangle$. This defines a maximally random ensemble of graphs with fixed expected values of the adjacency matrix elements. The Hamiltonian is given by

$$H = \sum_{i<j} \Theta_{ij}\sigma_{ij}. \tag{4.22}$$

The partition function is

$$Z = \sum_{G} e^{-H(G)} = \sum_{\sigma_{ij}} \exp\left(-\sum_{i<j}\Theta_{ij}\sigma_{ij}\right) = \prod_{i<j}\sum_{\sigma_{ij}=0}^{1} e^{-\Theta_{ij}\sigma_{ij}} = \prod_{i<j}\left(1+e^{-\Theta_{ij}}\right). \tag{4.23}$$

Thus the free enery in this case is given by

$$F = -\sum_{i<j}\ln\left(1+e^{-\Theta_{ij}}\right), \tag{4.24}$$

The probability of an edge between nodes $i, j$ is simply computed by the partial derivatives of the free energy

$$p_{ij} = \langle a_{ij} \rangle = \frac{\partial F}{\partial \Theta_{ij}} = \frac{1}{e^{\Theta_{ij}}+1}. \tag{4.25}$$

The probability $P(G)$ of a graph in this ensemble can be written as

$$P(G) = \frac{e^{-H(G)}}{Z} = \prod_{i<j} p_{ij}^{a_{ij}}(1-p_{ij})^{1-a_{ij}}. \tag{4.26}$$

### 4.3.2 Statistical mechanics of the function → structure approach to networks

We will use this framework of graph ensembles based on statistical mechanics to determine the average degree, degree distribution and the clustering coefficient of the networks coming out of our network navigation game. Apparently, we cannot compute the exact edges in a given constellation, however, we will show, that connection probabilities $p_{ij}$ between the nodes can be captured analytically. We will use these connection probabilities as observables to define our maximally random ensemble of graphs with Hamiltonian

$$H = \sum_{i<j} \Theta_{ij}\sigma_{ij}, \tag{4.27}$$

with

$$\Theta_{ij} = \ln\left(\frac{1-p_{ij}}{p_{ij}}\right). \tag{4.28}$$

Our connection probabilities will, of course, depend on the position of the nodes. The positions will be treated as hidden variables $\pi_i = (r_i, \phi_i)$ of each node, sampled from the distribution $\varrho(\pi)$. Then each pair of nodes will be connected with probability $p_{ij} = p(\pi_i, \pi_j)$. Thus the probability $P(G)$ of a graph in this ensemble can be written as

$$P(G) = \int p(G|\boldsymbol{\pi})\varrho(\boldsymbol{\pi}d\pi) = \int \prod_{i<j} p_{ij}^{a_{ij}}(1-p_{ij})^{1-a_{ij}} \prod_{i=1}^{N} \varrho(\pi_i)d\pi_i, \tag{4.29}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots \pi_N)$. This equilibrium graph ensemble is thus fully defined by two functions: the hidden variable probability density function $\varrho(\pi)$ and the connection probability function $p(\pi_i, \pi_j)$.

Most commonly, the edge probabilities of graph ensembles originate from structural peculiarities of networks, i.e., to enable analytical investigation of an ensemble matching an observable structural pattern. In the case of the function → structure approach, we originate the connection probabilities from the function of the graph, which is providing maximum navigability to the constituting nodes. The structural peculiarities of our networks will arise from fulfilling this navigation functionality at the lowest possible cost.

# Chapter 5

# Function-structure analysis of navigable networks

The Nash equilibrium of the network navigation game capturing the navigation function (Figure 4.2) is not necessarily unique. There can exist different networks minimizing the cost defined above. In what follows, among all the NNG equilibria, we always select the unique one that minimizes the sum of distances span by its edges, thus making the NNG Nash equilibrium network construction deterministic. However, there also exist certain edges, which we call frame edges, necessarily present in any Nash equilibrium. Edge $u \to v$ is a frame edge if $u$ is the closest node to $v$. In this case $u$ cannot navigate to $v$ through any other nodes since there is no one closer to $v$ than $u$self, so that $u$ must connect directly to $v$ to reach it, Figure 4.3. If at least one of such edges is absent, the network is not fully navigable. More exactly there is a well defined "frame topology" $G_{\text{frame}}$ with scale-free out-degree distribution which is present in *every* Nash equilibrium, or social optimum of the NNG ($G_{\text{frame}} \subset G(s^*)$) and other possible games having navigation as an incentive ($p_s = 1$). In other words the frame topology serves as a skeleton of any equilibrium topology emerging from navigational games. The frame topology is defined as:

**Definition 1** (Frame topology)**.** *Let* $G_{\text{frame}} = \bigcup_{u=0}^{N-1}(u \times g_u)$*, where* $g_u = \{v | v \notin s_u \Rightarrow c_u = \infty\}$*.*

Practically, the arc $(u, v)$ is contained in $G_{\text{frame}}$ if and only if the $d(u, v)$-disk centred at $v$ does not contain any node other than $u$ (see Figure 5.1). This means that $u$ cannot reach $v$ by greedy routing through any other nodes then $v$, and so it must create an arc towards $v$ to avoid of having infinite cost. Note that the in-degree of each node in $G_{\text{frame}}$ will be exactly one.

In any Nash equilibrium of this game, each node computes its optimal strategy independently of others. In game theory, such equilibria are called dominant strategy equilibria. Moreover, the equilibrium is also a social optimum since one cannot create a fully navigable network using fewer edges.

## 5.1 General formula for the connection probability

Here we cast the problem in statistical terms. We estimate the percentage of pairs of nodes located at a given distance that is connected in the NNG equilibrium. We call

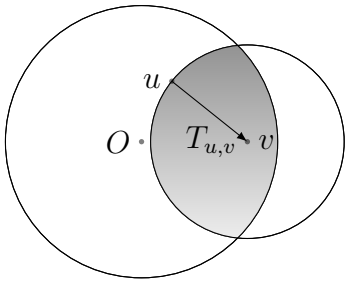



Figure 5.1: An edge in the $G_{\text{frame}}$

this percentage the effective connection probability. First, the connection probability of the Frame Topology is derived. This connection probability is a lower bound for the connection probability in the NNG equilibrium network because the Frame Topology is contained in every NNG equilibrium network. A direct upper bound of the connection probability is also studied. Based on a statistically equivalent lower bound and the direct upper bound, the general formula for the connection probability is induced, in which the average degree of the network is implicitly encoded. This makes it possible to approximate the connection probability (and all other quantities defined by it) using the observed average degree in the NNG simulation.

### 5.1.1 Connection probability in the Frame Topology

An arc $(u, v)$ in the Frame Topology is established if and only if there are no other nodes within the intersection of the $v-$centered disk with radius $d(u, v)$, and the original disk with radius $R$. Let $T_{uv}$ (see Figure 5.1) denote the area of the intersection of the $R$-disk and the $d(u, v)$-disk at origin $v$, and $\delta = N/T_R$ be the density of the points ($T_R$ denotes the area of the $R$-disk). The probability of this event is

$$\left(\frac{T_R - T_{uv}}{T_R}\right)^{N-2} \approx e^{-\delta T_{uv}} \tag{5.1}$$

An approximation for $T_{uv}$ is as follows: $T_{uv}$ is apparently equals to $2\pi(\cosh d_{uv} - 1)(\approx \pi e^{\frac{d_{uv}}{2}}$ for not so small $d_{uv})$ when the $d_{uv}-$disk is completely inside the $R-$disk. On contrary, if $R - r_v < d(u, v)$ (there is real intersection) then much less evidently $T_{uv}$ is approximately $T_{uv} \approx 4e^{\frac{d_{uv}}{2}} e^{\frac{R-r_v}{2}}$. In Figure 5.2 two characteristic cases are depicted when there is real intersection of the $d_{uv}$-disk and the $R$-disk. Let the polar coordinates of node $v$ be $(r_v, \phi_v)$, and of node $u$ be $(r_u, \phi_u)$. Let $\phi = |\phi_u - \phi_v|$. The area $T_{uv}$ is the function of $r_u$, $r_v$, $\phi$, and $R$, and can be calculated as the sum of the two circle sectors with angle $2\alpha$, radius $d_{uv}$ and angle $2\beta$ radius $R$, and minus the area of the two triangles with angles $\alpha, \beta, \gamma$. That is

$$T_{uv} = 2\beta\left(\cosh(R) - 1\right) + 2\alpha\left(\cosh(d_{uv}) - 1\right) - 2\left(\pi - \alpha - \beta - \gamma\right) \ . \tag{5.2}$$

where the angles and $d_{uv}$ are given by the hyperbolic law of cosines, however, here the following simpler approximations are used (which are accurate enough when $r_u$ and $d_{uv}$ appear in exponents):

$$d_{uv} \approx R + r_v + 2\ln\frac{\beta}{2} \Rightarrow \beta \approx 2e^{\frac{d_{uv}}{2} - \frac{R+r_v}{2}} \tag{5.3}$$

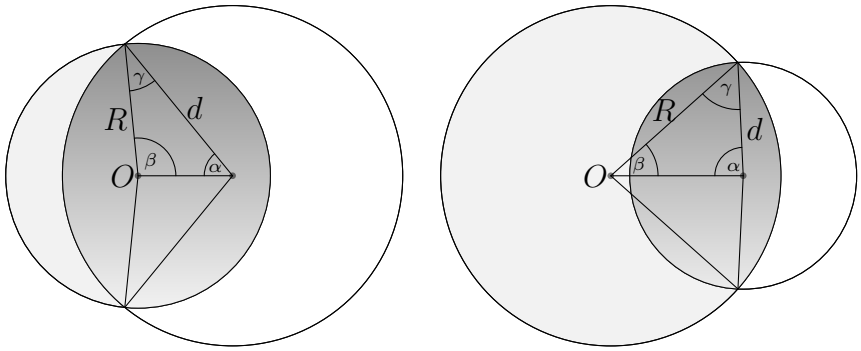



Figure 5.2: Illustration for $T_{uv}$

$$R \approx d_{uv} + r_v + 2\ln\frac{\alpha}{2} \Rightarrow \alpha \approx 2e^{-\frac{d_{uv}}{2}+\frac{R-r_v}{2}} \ . \tag{5.4}$$

Applying (5.2) with neglecting the triangle areas, and using $\cosh(R) - 1 \approx e^R/2$, $\cosh(d_{uv}) - 1 \approx \frac{e^{d_{uv}}}{2}$ we get $T_{uv} \approx 4e^{\frac{d_{uv}}{2}} e^{\frac{R-r_v}{2}}$.

In summary:

$$T_{uv} \approx \begin{cases} \pi e^{d_{uv}} & \text{, if } \ 0 < d_{uv} < R - r_v \\ 4e^{\frac{d_{uv}}{2}} e^{\frac{R-r_v}{2}} & \text{, if } \ d_{uv} > R - r_v \ . \end{cases} \tag{5.5}$$

This $T_{uv}$ approximations are illustrated in Figure 5.3 for $R = 12$, $r_v = 6$ and $r_v = 8$.

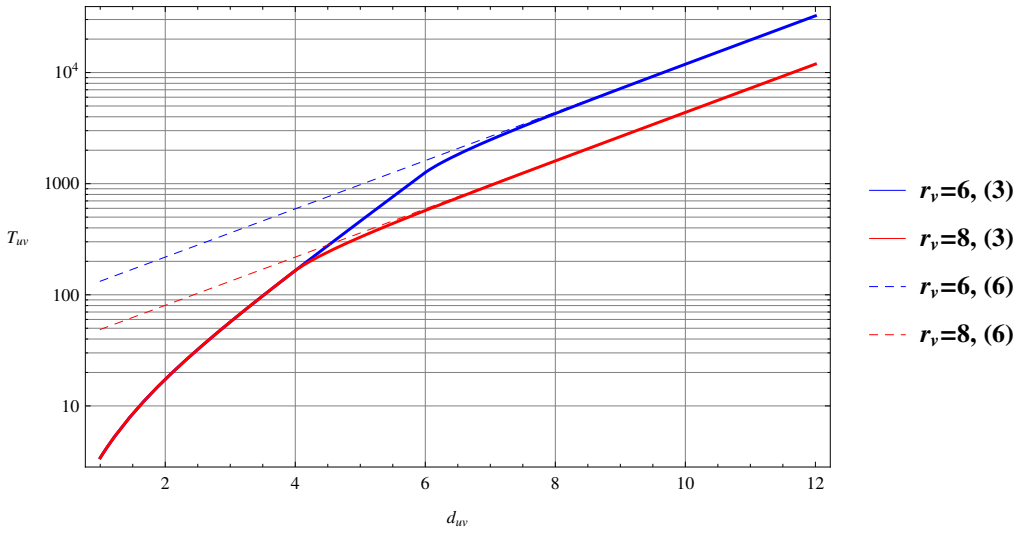



Figure 5.3: $T_{uv} \approx 4e^{\frac{d_{uv}}{2}} e^{\frac{R-r_v}{2}}$ when there are real intersections (that is when $d(u, v) > 6$ and 4, respectively).

Solid lines are the exact $T_{uv}$ calculations based on (5.2) and exact computations of angles. Note that there is a sharp change on logarithmic scale between the $d_{uv}$-slope and $d_{uv}/2$-slope around $R - r_v$. The dashed lines are the $T_{uv}$ approximations when $d_{uv} > R - r_v$.

The calculation of the expected degree of node $u$ requires $e^{-\delta T_{uv}}$ in the following double integration:

$$\delta \int_0^R \int_0^{2\pi} e^{-\delta T_{uv}} \mathrm{d}\phi \sinh(r_v) \mathrm{d}r_v \ . \tag{5.6}$$

Because the joint expansion of the double integral with respect to $r_v$ and $\phi$ reveals that the dominant terms will be those in which $d_{uv} > R - r_v$

$$\delta \int_0^R \int_0^{2\pi} e^{-\delta T_{uv}} \mathrm{d}\phi \sinh(r_v) \mathrm{d}r_v \approx \delta \int_0^R \int_0^{2\pi} e^{-\delta 4 e^{\frac{d_{uv}}{2}} e^{\frac{R-r_v}{2}}} \mathrm{d}\phi \sinh(r_v) \mathrm{d}r_v \ . \tag{5.7}$$

Using (5.3) it can also be shown that

$$\delta \int_0^R \int_0^{2\pi} e^{-\delta 4 e^{\frac{d_{uv}}{2}} e^{\frac{R-r_v}{2}}} \mathrm{d}\phi \sinh(r_v) \mathrm{d}r_v \approx \delta \int_0^R \int_0^{2\pi} e^{-\delta 8 e^{\frac{d_{uv}}{2}}} \mathrm{d}\phi \sinh(r_v) \mathrm{d}r_v \ . \tag{5.8}$$

Therefore,

$$\check{p}(d_{uv}) = e^{-\delta 8 e^{\frac{d_{uv}}{2}}} \tag{5.9}$$

can be considered as a statistically equivalent connection probability of the Frame Topology and as a latent (a statistically equivalent) lower bound of the connection probability of the equilibrium network of the NNG.

### 5.1.2 A direct upper bound for the connection probability

An upper bound for connection probability $p(d_{uv})$ can be derived as follows. Let $u$ and $v$ be two points in the $R$-disk and let $C_{u,v} = \{w | d_{wv} < d_{wu}\}$ denote the area for which $v$ is a good greedy next hop for $u$. This area is on the side of $v$ bounded by the perpendicular bisector ($B_{uv}$) of $(u, v)$, see Figure 5.4. (The figure is in the Poincare model). Let $A = \{x | C_{u,x} \supset C_{u,v}\}$. If there is a node $w \in A$ then

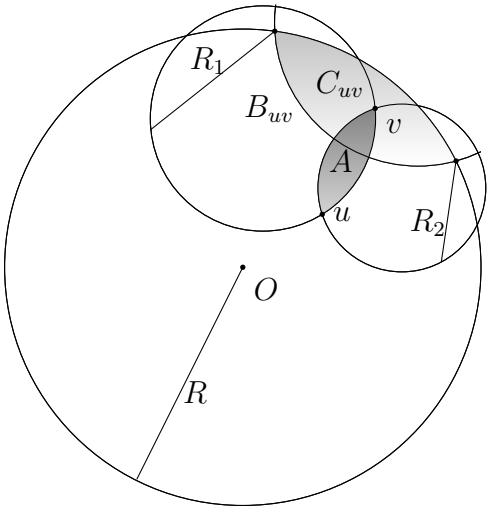



Figure 5.4: Calculation of $p(d_{uv})$

$u$ does not connect to $v$ since it has a node $w$ which covers the whole area that

$v$ can and some extra portion of the disk. Putting it differently $w$ can be in the optimal set cover (for $u$) instead of $v$. It is easy to see that $A$ is the intersection of two disks with radii $R_1$ and $R_2$ (the smaller circles on Figure 5.4). We can approximate the area of this intersection by the union of two sectors having angles $\phi_1 \approx 2e^{\frac{d_{uv}}{2}-R_1}$ and $\phi_2 \approx 2e^{\frac{d_{uv}}{2}-R_2}$ (by using an approximation on the hyperbolic distance $d_{uv} \approx 2R_i + 2\ln\frac{\phi_i}{2}$) of the $R_1$ and the $R_2$ disks respectively. Using this the area of $A$ is given by:

$$T_A \approx \phi_1(\cosh(R_1)-1) + \phi_2(\cosh(R_2)-1) \approx 2e^{\frac{d_{uv}}{2}-R_1}\frac{e^{R_1}}{2} + 2e^{\frac{d_{uv}}{2}-R_2}\frac{e^{R_2}}{2}, \quad (5.10)$$

which further simplifies to:

$$T_A \approx 2e^{\frac{d_{uv}}{2}}. \quad (5.11)$$

The probability that there is a node in $A$ is:

$$p(\exists w \in A) = 1 - \left(\frac{T_{\text{disk}} - T_A}{T_{\text{disk}}}\right)^{N-2} \approx 1 - e^{-\delta T_A}, \quad (5.12)$$

where $N$ denotes the number of nodes and $T_{\text{disk}}$ is the area of the $R$-disk. Trivially $p(d) \leq 1 - p(\exists w \in A)$ so:

$$p(d_{uv}) \leq e^{-\delta T_A}. \quad (5.13)$$

By substituting $T_A$ we get the following upper bound for the connection probability:

$$p(d_{uv}) \leq e^{-\delta T_A} \approx e^{-2\delta e^{\frac{d_{uv}}{2}}} =: \hat{p}(d_{uv}) \quad (5.14)$$

### 5.1.3 A general formula for the connection probability

In the Frame Topology (by definition), every node has exactly one incoming link; hence, the total number of links is $N$. From this, it immediately follows that the average out-degree of Frame Topology is 1. Regarding the direct upper bound of the connection probability, consider a network in which links are established by this upper bound probability. Also, the analysis in Section 5.2.1 implies that the average degree of such a network is 4. Based on the upper (5.14) and lower (5.9) bounds and the corresponding average degrees 1 and 4, a general formula of the connection probability can be induced as

$$p(d_{uv}, \delta, \bar{k}) = \text{Exp}\left(-\frac{8}{\bar{k}}\delta e^{\frac{d_{uv}}{2}}\right). \quad (5.15)$$

It will be shown in the next sections that a network formed by this connection probability has an average degree $\bar{k}$.

This formula is important because if an empirical average degree (which happens to be 2.27) can be observed in experiments (simulations) resulting in equilibrium networks of NNG, then not only upper and lower bounds on the expected degree of a node $u$ and degree distribution, but analytical approximations of them can also be given with this empirical mean. Figure 5.5 illustrates the relation of the upper and lower bounds, and the approximation of the connection probability to that of simulated NNG.

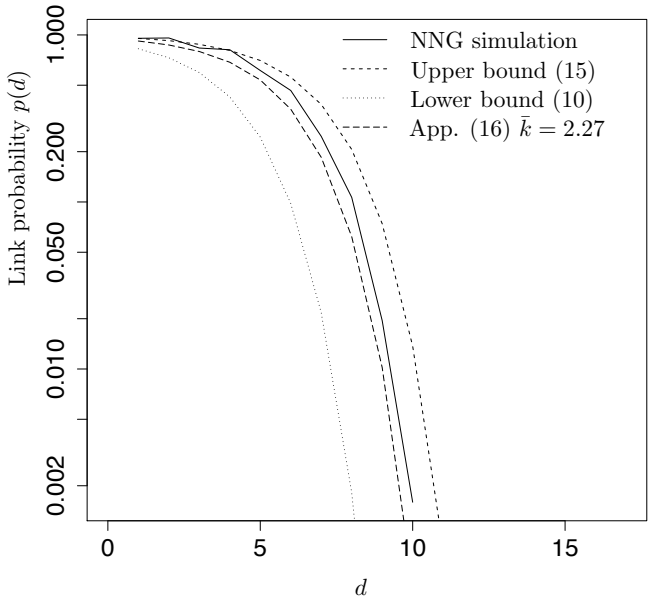



Figure 5.5: Connection probability as a function of distance in NNG simulations. The figure shows the analytic upper and lower bounds, as well as the analytic approximation with the empirical mean.

## 5.2 Structural properties of Nash-equilibrium networks.

Thus, if the node density is uniform ($\alpha = 1$), then the probability $p(d)$ that two nodes $u$ and $v$ located at distance $d \equiv d(u, v)$ are connected in a Nash equilibrium network lies between $\exp(-8\,\delta\,e^{d/2})$ and $\exp(-2\,\delta\,e^{d/2})$,

$$e^{-8\,\delta\,e^{d/2}} \leq p(d) \leq e^{-2\,\delta\,e^{d/2}} \ , \tag{5.16}$$

where $\delta$ is the average density of nodes on the disk, that is $\delta = N/A$, where $A$ is the disk area.

### 5.2.1 Expected degree

The expected degree of node $u$ at polar coordinates $(r_u, 0)$ — we can assume that $u$'s angular coordinate is $\phi_u = 0$ without loss of generality — is then $\bar{k}(r_u) = N \int p[d(u,v)]\,\rho(r_v)\rho(\phi_v)\,dr_v\,d\phi_v$, where $\rho(r_v)$ and $\rho(\phi_v)$ are the node densities from Eq. (4.1). The expected (out-) degree of a node $u$ with radial coordinate $r_u$ in a network generated with the effective connection probability formula is given by the following double integral

$$k_{\text{out}}(r_u, \bar{k}, R) = \delta \int_0^R \int_0^{2\pi} p(d_{uv}, \delta, \bar{k}) \mathrm{d}\phi \sinh(r_v) \mathrm{d}r_v \ . \tag{5.17}$$

The expected node-degree of the equilibrium network of NNG is lower bounded by $k_{\text{out}}(r_u, 1, R)$ (which coincides the expected node-degree of the Frame Topology)

whilst $k_{\text{out}}(r_u, 4, R)$ is the upper bound. An analytical approximation with the empirical mean $\bar{k} = 2.27$ can be given by $k_{\text{out}}(r_u, 2.27, R)$.

In what follows a formula is derived for $k_{\text{out}}(r_u, \bar{k}, R)$ based on the integral above. Considering the first integral by $\phi$ and applying the approximation (consider the hyperbolic law of cosine for $d_{uv}$, $r_u$, $r_v$, $\cosh d_{uv} = \cosh r_u \cosh r_v - \sinh r_u \sinh r_v \cos\phi$ )

$$e^{\frac{d_{uv}}{2}} \approx e^{\frac{r_u+r_v}{2}} \sqrt{\frac{1-\cos\phi}{2}} \tag{5.18}$$

we get that the integral can be approximated as

$$\delta \int_0^{2\pi} \text{Exp}\left(-\delta \frac{8}{\bar{k}} e^{\frac{d_{uv}}{2}}\right) \mathrm{d}\phi \approx 2\pi\delta(\mathrm{I}(0,x) - \mathrm{S}(0,x)) \approx \frac{1}{2}\bar{k} e^{-\frac{r_u+r_v}{2}} \tag{5.19}$$

where $x = \frac{8}{\bar{k}}\delta e^{\frac{r_u+r_v}{2}}$ and the last wave due to that $\mathrm{I}(0,x) - \mathrm{S}(0,x)$ (difference of the BesselI and the modified Struve functions) quickly tends to $\frac{2}{\pi}x^{-1}$ as $x$ increases [1]. Now the second integration by $r_v$ gives the expected degree approximation, that is

$$k_{\text{out}}(r_u, \bar{k}, R) \approx \int_0^R \frac{1}{2}\bar{k} e^{-\frac{r_u+r_v}{2}} \sinh(r_v) \mathrm{d}r_v \approx \frac{1}{2}\bar{k} e^{\frac{R}{2}} e^{-\frac{r_u}{2}} \ . \tag{5.20}$$

One can check that the average degree is indeed $\bar{k}$ with this expected node-degree:

$$\int_{r_u=0}^R \frac{1}{2}\bar{k} e^{\frac{R}{2}} e^{-\frac{r_u}{2}} \frac{\sinh r_u}{\cosh R - 1} \mathrm{d}r_u \approx \frac{1}{6}\bar{k}\,\mathrm{sech}^2\left(\frac{R}{4}\right)\left(\sinh\left(\frac{R}{2}\right) + 2\cosh\left(\frac{R}{2}\right) + 1\right) \approx \bar{k}\,. \tag{5.21}$$

We have numerically studied the accuracy of the approximations above. We have found that the exponential decay of the expected degree of nodes ($k_{\text{out}}(r_u)$) is a good approximation of the numerically evaluated expected degree function for a wide range of node density $\delta \in [10^{-8}, 10^{-2}]$. For example, consider a Frame Topology ($\bar{k} = 1$) with $R = 16.5$, $n = 10000$. In this case $\delta = 2.17 \cdot 10^{-4}$. Figure 5.6 shows how the expected degree decay is matching the exponential decay. We observe that while at smaller $r_u$ there are some approximation errors, for larger values of $r_u$ the match is very good. To quantify further, we note that 99.9% of points have $r_u > 10$ (that is in case of uniformly distributed points on the R(=16.5)-disk, expectedly only about 10 points of the 10000 are inside the disk with radius 10). If we consider the relative errors of the matching one can reveal that for $r_u > 10$ it is smaller than 0.15%, that is for 99.9% of points the expected degree approximation has smaller than 0.15% relative error. To increase the number of points to $n = 30000$ and $n = 50000$ ($\delta = 6.54 \cdot 10^{-4}$, $\delta = 1.08 \cdot 10^{-3}$), the relative error is increasing, especially for smaller values of $r_u$, but still for 99.9% of the points the relative error smaller than 0.25% and 1%, respectively. If we dramatically decrease the node-density, for example $n = 500$, the relative errors also increase (compared to the $n = 10000$ case), however, it still remains under 0.2% for 99.9% of the points.

### 5.2.2 Degree distribution

So we find that the expected number $\bar{k}(r)$ of connections of a node at radial coordinate $r$ is bounded by

$$\frac{1}{2} e^{(R-r)/2} \le \bar{k}(r) \le 2\, e^{(R-r)/2} \ , \tag{5.22}$$

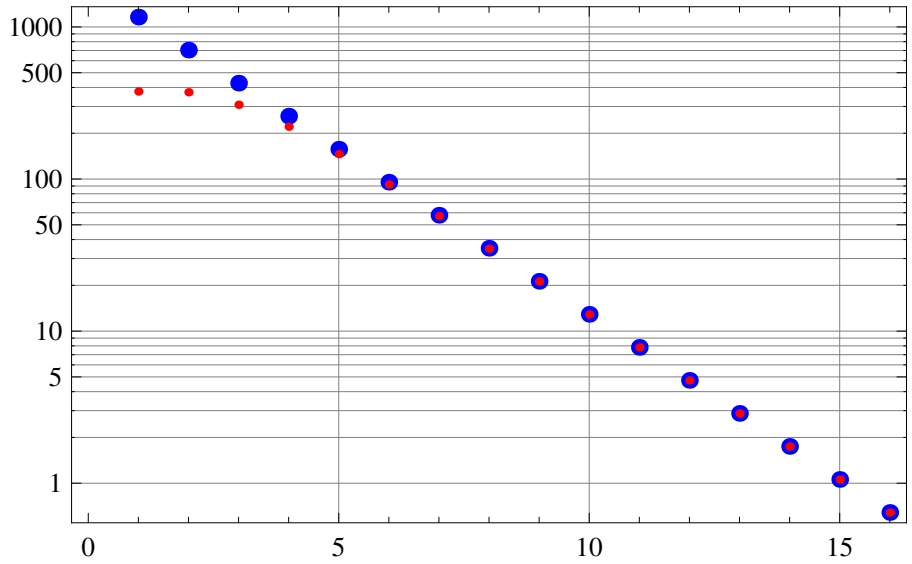



Figure 5.6: Exponential decay ($C(R)e^{-\frac{r_u}{2}}$, larger blue dots) versus the numerically evaluated exact decay (smaller red dots) of $u$'s expected degree as a function of $r_u$ ($R = 16.5$, $n = 10000$).

where $r \equiv r_u$. It then follows that the average degree of nodes in the network, given by $\bar{k} = \int_0^R \bar{k}(r)\rho(r)\,dr$, lies between 1 and 4,

$$1 \leq \bar{k} \leq 4. \tag{5.23}$$

We also see from Eq. (5.22) that the degree of nodes decays exponentially as the function of their radial position, $\bar{k}(r) \sim e^{-r/2}$, while their density exponentially increases, $\rho(r) \sim e^r$, Eq. (4.1). The combination of these two exponentials yields the power-law degree distribution in the network [28, 130]. Let us recall that in case of uniform distribution of points on an $R-$disk of the hyperbolic plane, the density of the radial coordinates of the points is

$$\rho(r) = \frac{\sinh r}{\cosh R - 1} \tag{5.24}$$

Note that the expected degree of node $u$ is exponential in the radial coordinate $r_u$ as in [139]. Because of this and the fact that equilibrium network of NNG is also sparse [24] the degree distribution can be calculated in the same way as in [139] :

$$P(k) = \int_0^R g(k, k_{\text{out}}(r_u))\rho(r_u)\mathrm{d}r_u = \frac{\bar{k}}{2}\frac{\Gamma(k-2, \frac{\bar{k}}{2})}{k!} \tag{5.25}$$

where $g(k, k_{\text{out}}(r_u))$ is the conditional distribution of the degree of a node with radial coordinate $u$, and it is Poissonian with mean $k_{\text{out}}(r_u)$ in case of sparse networks. It can also be shown that for larger $k$

$$P(k) \approx \frac{\bar{k}^2}{2k^3}\ . \tag{5.26}$$

The direct derivation of the complement cumulant degree distribution from $P(k)$ seems to be intangible, however, from its approximation it can be computed as

$$\bar{F}(k,\bar{k}) \approx 1 - \left( \int \frac{\bar{k}^2}{2k^3} \mathrm{d}k + C \right) \tag{5.27}$$

where the constant $C$ is 1, and $k \geq \frac{1}{2}\bar{k}$ (in order to have distribution function), that is

$$\bar{F}(k,\bar{k}) \approx \frac{\bar{k}^2}{4} k^{-2} \ , \quad k \geq \frac{1}{2}\bar{k} \ . \tag{5.28}$$

It is interesting to show that this approximation can also be obtained as the *exact* ccdf of the conditional expected node degrees $k_{\mathrm{out}}(r_u)$. This approximation can be computed as

$$\bar{F}(k,\bar{k}) \approx \int_{r=0}^{r_u(k)} \rho(r) \mathrm{d}r \approx e^{r_u(k)-R} \tag{5.29}$$

where $r_u(k)$ is the inverse function of $k_{\mathrm{out}}(r_u,\bar{k},R)$ w.r.t. $r_u$, i.e.

$$r_u(k) = R - 2\ln\left(2k/\bar{k}\right) \ . \tag{5.30}$$

Applying this one can obtain the same before as

$$\bar{F}(k,\bar{k}) \approx \frac{\bar{k}^2}{4} k^{-2} \ , \quad k \geq \frac{1}{2}\bar{k} \ . \tag{5.31}$$

Note that this yields the average degree equal to $\bar{k}$ as expected:

$$\int_{k=\frac{1}{2}\bar{k}}^{\infty} \left( k \frac{\partial(1-\bar{F}(k,\bar{k}))}{\partial k} \right) = \bar{k}. \tag{5.32}$$

From this, an analytical approximation of the ccdf of the NNG equilibrium network is $\bar{F}(k,2.27)$, its lower and upper bounds are $\bar{F}(k,1)$, $\bar{F}(k,4)$, respectively. In Figure 5.7 these analytical formulae are also drawn with a completely empirical distribution obtained from NNG simulation.

We also note that the $\delta$-independence of $k_{\mathrm{out}}(r_u)$ and $\bar{F}(k)$ is approximate, but it holds with a high accuracy for $\delta \in [10^{-8}, 10^{-2}]$, including the frame topology.

So the degree distribution is:

$$P(k) = \frac{1}{k!} \int_0^R e^{-\bar{k}(r)} [\bar{k}(r)]^k \rho(r)\, dr = 2 \left(\frac{\bar{k}}{2}\right)^2 \frac{\Gamma(k-2,\bar{k}/2)}{k!} \sim k^{-3}. \tag{5.33}$$

.

### 5.2.3 Clustering coefficient

Here we analyze local clustering using the effective connection probability (5.15). By means of quasi-symbolic calculations we also show that local clustering depends on the expected node degree $k$ similarly for both lower and upper bounds of the effective connection probability, and that average clustering does not depend on average degree $\bar{k}$.

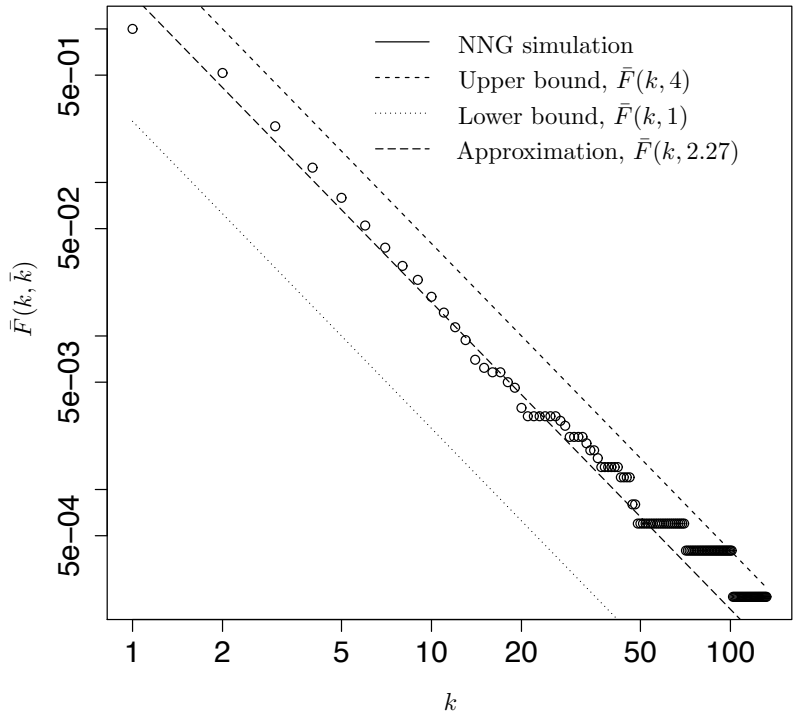



Figure 5.7: Empirical CCDF of the degree distribution, its analytical upper and lower bounds $\bar{F}(k,4)$, $\bar{F}(k,1)$, and analytical approximation with the empirical mean $\bar{F}(k,2.27)$.

Let the hyperbolic polar coordinates of the point triplet $u,v,w$ be $(r_u,\phi_u),(r_v,\phi_v),(r_w,\phi_w)$ and $\phi=\phi_u-\phi_v$, $\psi=\phi_u-\phi_w$. The local clustering coefficient $cl(r_u)$ for a given node $u$ is calculated as the ratio of the expected number of link pairs with common edge $u$ and the expected number of link triangles with edge $u$. For calculating these expected numbers, the joint probabilities of the existence of $(u,v)$ and $(u,w)$ link pair and the existence of the $(u,v,w)$ link triangle are substituted by $p(d_{uv})p(d_{uw})$ and $p(d_{uv})p(d_{uw})p(d_{vw})$, respectively. This requires link independence assumption, which is not true, however, correlations are expectedly diminished due to averaging processes (like in mean field calculations [66]). In this way, the local clustering coefficient is formulated as

$$cl(r_u)=\frac{\delta^2\int_{r_w=0}^{R}\int_{r_v=0}^{R}\int_{\psi=0}^{2\pi}\int_{\phi=0}^{2\pi}p(d_{uv})p(d_{uw})p(d_{vw})\mathrm{d}\phi\mathrm{d}\psi\sinh(r_v)\sinh(r_w)\mathrm{d}r_v\mathrm{d}r_w}{\delta^2\int_{r_w=0}^{R}\int_{r_v=0}^{R}\int_{\psi=0}^{2\pi}\int_{\phi=0}^{2\pi}p(d_{uv})p(d_{uw})\mathrm{d}\phi\mathrm{d}\psi\sinh(r_v)\sinh(r_w)\mathrm{d}r_v\mathrm{d}r_w}\,. \tag{5.34}$$

For estimating these integrals in the numerator and the denominator the following functions are defined:

$$\int_{\psi=0}^{2\pi}\int_{\phi=0}^{2\pi}p(d_{uv})p(d_{uw})p(d_{vw})\mathrm{d}\phi\mathrm{d}\psi\approx \tag{5.35}$$

$$\approx\int_{\psi=0}^{2\pi}\int_{\phi=0}^{2\pi}\exp\left(-x\sin\frac{\phi}{2}-y\sin\frac{\psi}{2}-z\sin\frac{|\psi-\phi|}{2}\right)\mathrm{d}\phi\mathrm{d}\psi=:\mathrm{Nu}(x,y,z)$$

and

$$\int_{\psi=0}^{2\pi}\int_{\phi=0}^{2\pi} p(d_{uv})p(d_{uw})\mathrm{d}\phi\mathrm{d}\psi \approx \int_{\psi=0}^{2\pi}\int_{\phi=0}^{2\pi} \exp\left(-x\sin\frac{\phi}{2} - y\sin\frac{\psi}{2}\right)\mathrm{d}\phi\mathrm{d}\psi =: \mathrm{De}(x,y) \tag{5.36}$$

where the general connection probability formula (5.15), the approximation $e^{\frac{d_{uv}}{2}} \approx e^{\frac{r_u+r_v}{2}}\sqrt{\frac{1-\cos\phi}{2}}$ are applied and

$$x = \frac{8}{\bar{k}}\delta e^{\frac{r_u+r_v}{2}}\ ,\ y = \frac{8}{\bar{k}}\delta e^{\frac{r_u+r_w}{2}}\ ,\ z = \frac{8}{\bar{k}}\delta e^{\frac{r_v+r_w}{2}}\ . \tag{5.37}$$

Now we apply asymptotic expansions of $\mathrm{Nu}(x,y,z)$ and $\mathrm{De}(x,y)$ in order to approximate them. (Asymptotic expansion here means that $x,y,z$ are large parameters and we are interested in the asymptotic behaviour of these integrals as $\{x,y,z\} \to \infty$). Note that $\mathrm{De}(x,y)$ is simply the product of two integrals which reads as

$$\begin{aligned}\mathrm{De}(x,y) &:= \int_{\psi=0}^{2\pi}\exp\left(-y\sin\frac{\psi}{2}\right)\mathrm{d}\psi\int_{\phi=0}^{2\pi}\exp\left(-x\sin\frac{\phi}{2}\right)\mathrm{d}\phi \\ &= 4\pi^2(\mathrm{I}(0,x)-\mathrm{S}(0,x))(\mathrm{I}(0,y)-\mathrm{S}(0,y)) \approx \frac{16}{xy}\end{aligned} \tag{5.38}$$

due to that $\mathrm{I}(0,x)$-$\mathrm{S}(0,x) \approx \frac{2}{\pi}x^{-1}$ based on its asymptotic expansion [1] .

For approximating $\mathrm{Nu}(x,y,z)$ we use Laplace's [22] method to generate first orders of the asymptotic expansion with respect to $x,y$, and $z$. For this, we take the first-order Taylor series expansion of the sinus functions around 0 and $2\pi$ where the integral is dominant for larger $x,y,z$. Performing the double integral (5.35) with these series and erasing the exponentially small terms, we get the following four terms with respect to that $x$ is in the neighborhood of 0 or $2\pi$ and $y$ is in the neighborhood of 0 or $2\pi$ :

$$\mathrm{Nu}(x,y,z) \approx 2\frac{4(x+y+2z)}{(x+y)(x+z)(y+z)} + 2\frac{4}{(x+z)(y+z)} = \frac{16(x+y+z)}{(x+y)(x+z)(y+z)}\ . \tag{5.39}$$

Now the clustering coefficient can be written as

$$\begin{aligned}cl(r_u) &\approx \frac{\frac{\delta^2}{2}\int_{r_w=0}^{R}\int_{r_v=0}^{R}\mathrm{Nu}(x,y,z)\sinh(r_v)\sinh(r_w)\mathrm{d}r_v\mathrm{d}r_w}{\frac{\delta^2}{2}\int_{r_w=0}^{R}\int_{r_v=0}^{R}\mathrm{De}(x,y,z)\sinh(r_v)\sinh(r_w)\mathrm{d}r_v\mathrm{d}r_w} \approx \\ &\approx \frac{\int_{r_w=0}^{R}\int_{r_v=0}^{R}\frac{16(x+y+z)}{(x+y)(x+z)(y+z)}\sinh(r_v)\sinh(r_w)\mathrm{d}r_v\mathrm{d}r_w}{\int_{r_w=0}^{R}\int_{r_v=0}^{R}\frac{16}{xy}\sinh(r_v)\sinh(r_w)\mathrm{d}r_v\mathrm{d}r_w}\end{aligned} \tag{5.40}$$

Based on this it can be seen that $cl(r_u)$ does NOT depend on the density parameter $\delta$, and depends on the average degree $\bar{k}$ only through $r_u(k,\bar{k})$ (see equation (5.20) ) because all the $x,y,z$ terms contain a $\frac{8}{k}\delta$ factor. In this way both integrals in the numerator and denominator posses a $\frac{1}{\delta^2}$ factor. (Note, that both the numerator and denominator are independent from $\delta$).

In what follows we explore how the local clustering coefficient of a node is depending on the expected degree $k$. This is possible to perform through the inverse function of $\bar{k}(r_u)$ (based on (5.20)) which is $r_u(k) = R - 2\ln\left(2k/\bar{k}\right)$ . First the denominator is calculated which is possible in a parametric way.

$$\int_{r_w=0}^{R}\int_{r_v=0}^{R}\frac{16}{xy}\sinh(r_v)\sinh(r_w)\mathrm{d}r_v\mathrm{d}r_w = \frac{1}{9}e^{-4R}(1-4e^{3R/2}+3e^{2R})^2k^2 \approx k^2 \quad (5.41)$$

with the substitutions $x, y$ in (5.37) and $r_u(k)$ above. (The term $\frac{16}{xy}$ does not depend on $\bar{k}$ due to the $x, y$ and $r_u(k)$ substitution). Note that this is a good cross-validation of this formula, because the expected number of link pairs of a node with given expected degree $k$ is approximately $k(k-1)/2 \approx k^2/2$. This is because if the node degree $\kappa$ has Poisson distribution with parameter $k$ then the expected number of link pairs at this node is $E\left[\frac{\kappa(\kappa-1)}{2}\right] = \sum_{l=0}^{\infty}\frac{l(l-1)}{2}\frac{k^l}{l!}e^{-l}$, which is exactly $\frac{k^2}{2}$. Based on the equations (5.40), (5.41) and substituting $x, y, z$ into the formula of the integrand one can obtain

$$cl(k,\bar{k},R) \approx \int_{r_w=0}^{R}\int_{r_v=0}^{R}\frac{\bar{k}e^{\frac{1}{2}(r_v+r_w-R)}\left(\bar{k}e^{\frac{1}{2}(r_v+R)}+\bar{k}e^{\frac{1}{2}(r_w+R)}+2ke^{\frac{1}{2}(r_v+r_w)}\right)}{4\left(e^{\frac{r_v}{2}}+e^{\frac{r_w}{2}}\right)\left(e^{R/2}\bar{k}+2ke^{\frac{r_v}{2}}\right)\left(e^{R/2}\bar{k}+2ke^{\frac{r_w}{2}}\right)}\mathrm{d}r_v\mathrm{d}r_w\,. \quad (5.42)$$

This double integral on the right hand side can be assessed symbolically by substitution, but even a simplified result is still quite spacious. Nevertheless, the detailed analysis of this function reveals that it is approximately independent of $R$, and as $k$ is increasing, the local clustering coefficient tends to

$$cl(k,\bar{k}) \approx \ln(2)\bar{k}k^{-1}\,. \quad (5.43)$$

For simplicity and for catching the behaviour of $cl(k,\bar{k})$ even for smaller $k$ values, the following intuitive form of approximation is calculated by numerical matching. The intuition is based on the observation that the integrand itself is in the form of a fraction of a first order and a second order polynomial of $k$.

$$\int_{r_w=0}^{R}\int_{r_v=0}^{R}\frac{\bar{k}e^{\frac{1}{2}(r_v+r_w-R)}\left(\bar{k}e^{\frac{1}{2}(r_v+R)}+\bar{k}e^{\frac{1}{2}(r_w+R)}+2ke^{\frac{1}{2}(r_v+r_w)}\right)}{4\left(e^{\frac{r_v}{2}}+e^{\frac{r_w}{2}}\right)\left(e^{R/2}\bar{k}+2ke^{\frac{r_v}{2}}\right)\left(e^{R/2}\bar{k}+2ke^{\frac{r_w}{2}}\right)}\mathrm{d}r_v\mathrm{d}r_w \approx \frac{1+ak}{b+ck+dk^2} \quad (5.44)$$

where the coefficient $a, b, c, d$ are approximately independent of $R$ and is depending only on $\bar{k}$. The coefficient is summarised in Table 5.1 for three cases: for the lower bound of the average degree 1, for the upper bound 4, and $\bar{k} = 2.27$ which latter average degree comes from the numerical simulation of the network formation game.

Note that, for larger $k$'s

$$\frac{1+ak}{b+ck+dk^2} \approx \frac{a}{d}k^{-1}\,,\ \frac{1}{2}\bar{k} \le k \le \frac{1}{2}\bar{k}e^{\frac{R}{2}}\,, \quad (5.45)$$

and $\frac{a}{d}$ is very close to $\ln(2)\bar{k}$ for all the three cases, as expected.

It is now possible to compute average clustering based on the approximation above as

$$cl = \int_{k=\frac{1}{2}\bar{k}}^{\frac{1}{2}\bar{k}e^{\frac{R}{2}}} cl(k,\bar{k})\frac{\partial}{\partial k}(1-\bar{F}(k))\mathrm{d}k \approx \int_{k=\frac{1}{2}\bar{k}}^{\frac{1}{2}\bar{k}e^{\frac{R}{2}}}\frac{1+ak}{b+ck+dk^2}\frac{\bar{k}^2}{2k^3}\mathrm{d}k\,. \quad (5.46)$$

| $\bar{k}$ | a | b | c | d |
|---|---|---|---|---|
| 1 | 0.598 | 1.008 | 2.168 | 0.869 |
| 2.27 | 0.331 | 1.002 | 1.019 | 0.209 |
| 4 | 0.220 | 1.002 | 0.618 | 0.080 |

Table 5.1: The clustering coefficient as a function of the average degree.

Evaluating this integral for the average degree lower bound $\bar{k} = 1$, upper bound $\bar{k} = 4$, and the average degree in simulations $\bar{k} = 2.27$, we obtain, using Table 5.1, $cl = 0.447075, 0.447615, 0.447146$, respectively. We have also performed more extensive numerical experiments showing that average clustering does not significantly depend on the average degree for $\delta \in [10^{-8}, 10^{-2}]$, and $R \in [10, 20]$. Its dependence on $R$ is also negligible, which is not surprising since $R$ appears only on the upper limit of the integral, and this upper limit negligibly affects the result since the integrands decrease as $\sim k^{-5}$. All these analytic and numeric results are in a good agreement with simulations, see Figure 5.8.

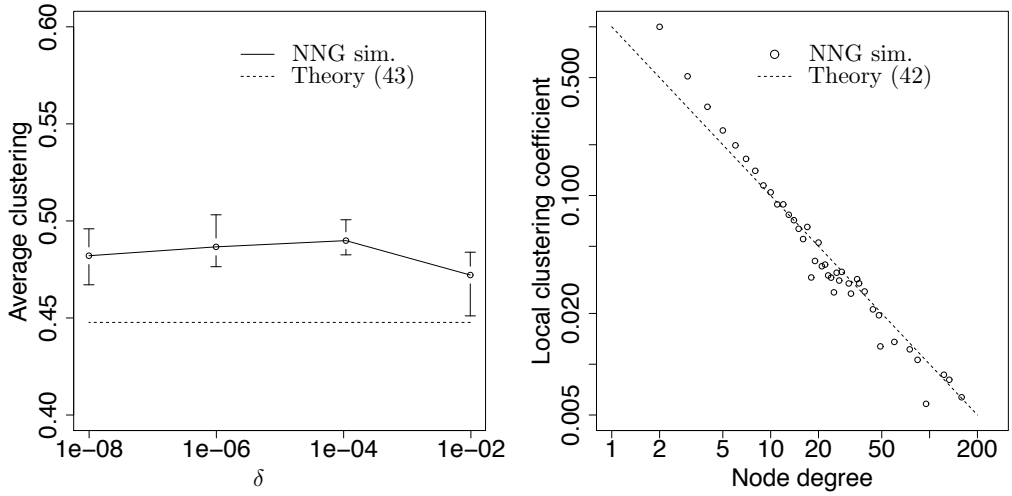



Figure 5.8: Average clustering as a function of $\delta$, and local clustering as a function of node degree.

**Evaluating the integral**

The integral for computing the local clustering coefficient presented (5.42) can be evaluated by the following substitution

$$\xi = \mathrm{Exp}\left(\frac{r_v}{2}\right) \ , \ \zeta = \mathrm{Exp}\left(\frac{r_w}{2}\right) \ , \ \mathrm{d}\xi = \mathrm{Exp}\left(\frac{r_v}{2}\right)\frac{1}{2}\mathrm{d}r_v \ , \ \mathrm{d}\zeta = \mathrm{Exp}\left(\frac{r_w}{2}\right)\frac{1}{2}\mathrm{d}r_w \ , \tag{5.47}$$

which is in the form

$$cl(k, \bar{k}, R) \approx \int_1^{e^{\frac{R}{2}}} \int_1^{e^{\frac{R}{2}}} \frac{e^{-\frac{R}{2}\bar{k}}\left(e^{\frac{R}{2}\bar{k}}(\zeta+\xi) + 2\zeta k\xi\right)}{(\zeta+\xi)\left(e^{\frac{R}{2}\bar{k}} + 2\zeta k\right)\left(e^{\frac{R}{2}\bar{k}} + 2k\xi\right)} \mathrm{d}\xi\mathrm{d}\zeta \ . \tag{5.48}$$

A simplified version of the result of the integral (5.42) is

$$\frac{1}{8k^2}\left(e^{-\frac{R}{2}}\left(\bar{k}e^{R/2}\left(\bar{k}\left(\mathrm{Li}_2\left(\frac{2\left(1+e^{-\frac{R}{2}}\right)k}{2k-\bar{k}}\right)-\mathrm{Li}_2\left(\frac{2\left(1+e^{-\frac{R}{2}}\right)k}{2k+\bar{k}}\right)+\right.\right.\right.\right. \tag{5.49}$$

$$\mathrm{Li}_2\left(-\frac{2\left(1+e^{R/2}\right)k}{e^{R/2}\bar{k}-2k}\right)+\mathrm{Li}_2\left(\frac{4k}{2k+e^{R/2}\bar{k}}\right)-\mathrm{Li}_2\left(\frac{2\left(1+e^{R/2}\right)k}{2k+e^{R/2}\bar{k}}\right)-\mathrm{Li}_2\left(\frac{4k}{2k-e^{R/2}\bar{k}}\right)-$$

$$\mathrm{Li}_2\left(-\frac{4k}{\bar{k}-2k}\right)+\mathrm{Li}_2\left(\frac{4k}{2k+\bar{k}}\right)\Bigg)+\bar{k}\left(\log\left(e^{R/2}+1\right)\left(\ln\left(-\frac{2ke^{-\frac{R}{2}}+\bar{k}}{2k-\bar{k}}\right)-\ln\left(\frac{\bar{k}-2ke^{-\frac{R}{2}}}{2k+\bar{k}}\right)+\right.\right.$$

$$\ln\left(\frac{e^{R/2}(2k+\bar{k})}{\bar{k}e^{R/2}-2k}\right)-\ln\left(\frac{e^{R/2}(\bar{k}-2k)}{2k+\bar{k}e^{R/2}}\right)\Bigg)+\ln\left(2e^{R/2}\right)\left(\ln\left(1-\frac{4k}{2k+\bar{k}}\right)-\ln\left(\frac{4k}{\bar{k}-2k}+1\right)\right)+$$

$$\ln(2)\left(\ln\left(1-\frac{4k}{2k+\bar{k}e^{R/2}}\right)-\ln\left(\frac{4k}{\bar{k}e^{R/2}-2k}+1\right)\right)+2\left(\ln\left(e^{R/2}(2k+\bar{k})\right)-\right.$$

$$\left.\ln\left(2k+\bar{k}e^{R/2}\right)\right)\left(\tanh^{-1}\left(\frac{2k}{\bar{k}}\right)-\tanh^{-1}\left(\frac{2ke^{-\frac{R}{2}}}{\bar{k}}\right)\right)\Bigg)+8k\left(\ln\left(e^{R/2}\right)-\ln\left(e^{R/2}+1\right)+\ln(2)\right)+$$

$$\left.\left.4k\bar{k}\left(\ln(4)-2\ln\left(e^{R/2}+1\right)\right)\right)\right),$$

where the function $\mathrm{Li}_2(z)=\sum_{k=1}^{\infty}\frac{z^k}{k^2}$ is the di-logarithm special function. We observe that factors $\mathrm{Exp}(-R/2)$ and $\mathrm{Exp}(R/2)$ appear in several terms. If $R$ is sufficiently large, e.g., ranging between realistic values of 10 and 20, then we can neglect the exponentially smaller terms, keeping only the exponentially large dominating terms. For example,

$$\frac{\bar{k}-2ke^{-\frac{R}{2}}}{2k+\bar{k}}\approx\frac{\bar{k}}{2k+\bar{k}}\quad\text{and}\quad\frac{e^{R/2}(2k+\bar{k})}{\bar{k}e^{R/2}-2k}\approx\frac{2k+\bar{k}}{\bar{k}}\ . \tag{5.50}$$

Using this procedure, after some simplifications, we finally obtain an $R-$free expression for clustering:

$$cl(k,\bar{k})\approx\frac{1}{8k^2}\bar{k}\left(8k\ln(2)+\bar{k}\left(\ln\left(\frac{\bar{k}+2k}{\bar{k}}\right)\ln\left(\frac{\bar{k}+2k}{\bar{k}-2k}\right)+\ln(2)\ln\left(\frac{\left(\bar{k}-2k\right)^2}{\left(\bar{k}+2k\right)^2}\right)\right)+\right.$$

$$\bar{k}\left(\mathrm{Li}_2\left(\frac{2k}{2k-\bar{k}}\right)+\mathrm{Li}_2\left(-\frac{2k}{\bar{k}}\right)-\mathrm{Li}_2\left(\frac{2k}{\bar{k}}\right)-\right.$$

$$\left.\left.\mathrm{Li}_2\left(-\frac{4k}{\bar{k}-2k}\right)-\mathrm{Li}_2\left(\frac{2k}{2k+\bar{k}}\right)+\mathrm{Li}_2\left(\frac{4k}{2k+\bar{k}}\right)\right)\right) \tag{5.51}$$

We can now see that $cl(k,\bar{k})\to\ln(2)\bar{k}\ k^{-1}$ as $k$ increases, because the logarithmic terms become zero, while the dilogarithmic terms eliminate each other. The analysis of this function at $k=0$ also shows that $cl(0,\bar{k})=1$, from which it follows that $b=1$ in the polynomial matching the numerical calculations, cf. Table 5.1.

In summary, the average clustering $\bar{c}(k)$ of nodes of degree $k$ decays with $k$ as $1/k$, while the average clustering $\bar{c}=\sum_k P(k)\bar{c}(k)$ in the network is around 0.45,

also confirmed in simulations. Clustering does not depend on network size or average degree, meaning that clustering is a positive constant even in the large graph size limit. Remarkably, neither degree distribution nor clustering depends on the node density $\delta$.

### 5.2.4 Non-uniform node density

For non-uniform node density $\alpha \neq 1$, we can analytically obtain only the lower bound for $\bar{k}(r,\alpha)$, which is still proportional $e^{-\frac{r}{2}}$, i.e., independent of $\alpha$ if $\alpha > 1/2$.

The radial coordinate density in case of quasi-uniform node density is

$$\rho(r,\alpha) := \frac{\alpha \sinh(\alpha r)}{\cosh(\alpha R) - 1} \approx \alpha e^{\alpha(r-R)} \tag{5.52}$$

while the angle density remains uniform ($\frac{1}{2\pi}$) over the range $[1, 2\pi]$. Given a point pair $(u,v)$, first we determine the probability $p(r_u,\alpha)$ that the $u \to v$ link exists, then based on this the average out degree $k(r_u,\alpha)$ of $u$ is calculated, and finally $\bar{F}(k,\alpha)$ is also given.

Probability $p(r_u,\alpha)$ is equal to the probability that none of the remaining $N-2$ points fall in the intersection of the $v$-centred $d_{uv}$ circle and the $R$-disk. Let us denote by $p_1$ the probability that a point whose coordinates generated randomly according to the densities above falls inside the intersection. Using $p_1$ the probability $p(r_u,\alpha)$ can be calculated and approximated as

$$p(r_u,\alpha) = (1-p_1)^{N-2} \approx e^{-Np_1} \tag{5.53}$$

The calculation of $p_1$ can be performed by using the node density function in the following way [139]

$$p_1 = \int_0^{\max(0,d-r_v)} \rho(r,\alpha)\mathrm{d}r + \frac{1}{2\pi}\int_{|d-r_v|}^{\min(R,d+r_v)} \rho(r,\alpha)2\theta(r)\mathrm{d}r \tag{5.54}$$

where

$$\theta(r) = \arccos\frac{\cosh r_v \cosh r - \cosh d}{\sinh r_v \sinh r}\ . \tag{5.55}$$

In [139] a useful approximation is presented for quite similar integrals, based on which one can write

$$p_1 \approx \frac{4e^{\frac{1}{2}(d-R-r_v)}\alpha}{\pi(-1+2\alpha)} \tag{5.56}$$

for $0.5 < \alpha \leq 1$ .

Now the expected out-degree of $u$ can be written as

$$k_{\mathrm{out}}(r_u,\alpha) \approx \frac{N}{2\pi}\int_0^R\int_0^{2\pi} e^{-Np_1}\mathrm{d}\phi\rho(r_v)\mathrm{d}r_v\ . \tag{5.57}$$

Using the approximation of $p_1$ and $\cosh(d/2) \approx e^{\frac{r_u+r_v}{2}}\sin\frac{\phi}{2}$ one can formulate

$$\int_0^{2\pi} e^{-Np_1}\mathrm{d}\phi \approx \int_0^{2\pi} e^{-x\sin\frac{\phi}{2}}\mathrm{d}\phi \approx 2\pi(\mathrm{I}(0,x) - \mathrm{S}(0,x)) \approx \frac{4}{x} \tag{5.58}$$

where

$$x = 4\frac{N}{\pi}\frac{\alpha}{2\alpha - 1}e^{\frac{r_u - R}{2}} \ . \tag{5.59}$$

Note, that $x$ does not depend on $r_v$, therefore the second integration by $r_v$ results

$$k_{\text{out}}(r_u, \alpha) \approx \frac{N}{2\pi}\frac{4}{x}\int_0^R \rho(r_v, \alpha)\mathrm{d}r_v = \frac{2\alpha - 1}{2\alpha}e^{\frac{R}{2}}e^{-\frac{r_u}{2}} \ . \tag{5.60}$$

Note, that for $\alpha = 1$ we get back the result for the uniform density case, (5.20).

Now the (approximation of the) complement cumulative distribution function $\bar{F}(k)$ can be derived as,

$$\bar{F}(k) = \int_0^{r_u(k)} \rho(r, \alpha)\mathrm{d}r \approx e^{\alpha(r_u(k) - R)} = \left(\frac{1 - \frac{1}{2\alpha}}{k}\right)^{2\alpha} \tag{5.61}$$

where $r_u(k)$ is the inverse function of $k_{\text{out}}(r_u)$ . The simulation results displayed in Figure 5.9 readily confirm this finding.

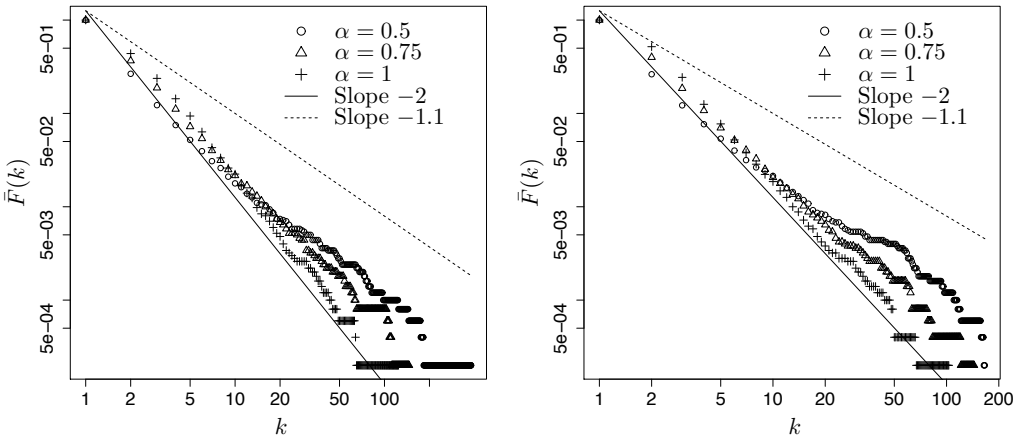
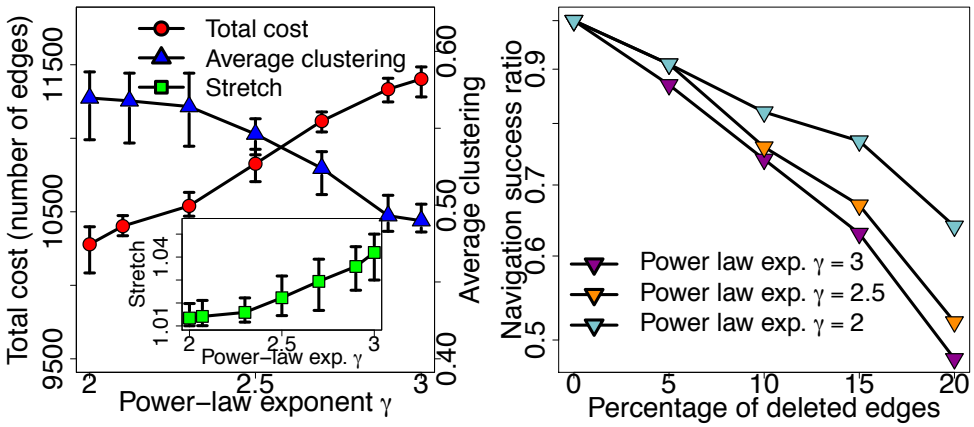



Figure 5.9: The in and out degree distributions of the NNG for various settings of the $\alpha$ parameter.

This lower bound suggests that the degree distribution is a power-law $P(k) \sim k^{-\gamma}$ with exponent $\gamma = 2\alpha + 1$. Figure 5.10 shows that the closer the $\gamma$ to 2, the stronger the clustering, the cheaper the network, and the more efficient and robust the navigability is. The value of $\gamma = 2$ thus appears as the "best choice" for a network—the network is maximally navigable at the lowest cost.

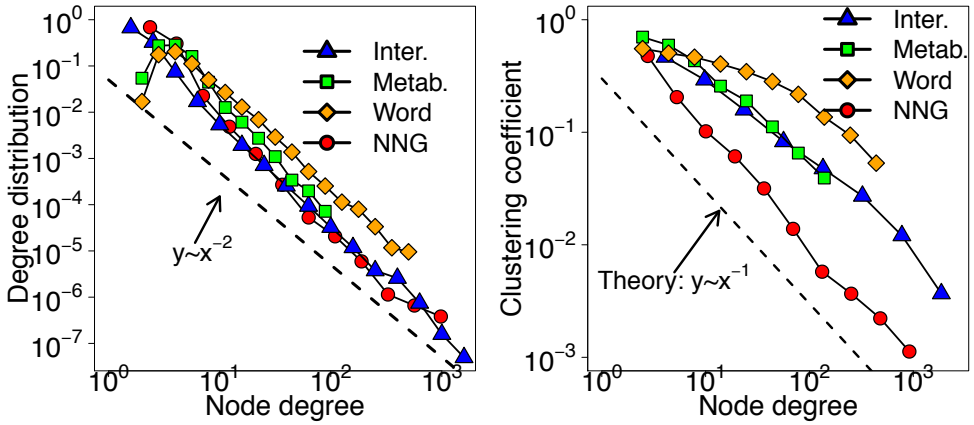
These results complement existing works [45, 26] showing that $\gamma = 2$ yields most navigable networks, by adding that this $\gamma$ provides a minimum cost equilibrium topology as well, explaining the emergence of these networks from the interaction of selfish players.

Figure 5.11 and Table 5.2 confirm our analytic results and shows that some basic structural properties of NNG-simulated networks are similar to some real networks. Our results also suggest that the incentive for navigability alone may be sufficient to explain the properties of complex networks to a certain degree. Yet we cannot really make this claim based only on such large-scale statistical similarities. A more detailed link-by-link comparison between real and corresponding NNG networks is needed to understand how well the NNG reflects reality.




(a) (b)

Figure 5.10: Topological properties of NNG equilibrium networks as a function of the power-law exponent. Panel (a) shows the total cost (number of edges), average clustering $\bar{c}$, and stretch in NNG-simulated networks as functions of $\gamma$. Stretch (shown in the inset) is the average factor showing by how much longer the greedy navigation paths are compared to the shortest paths in the network. Stretch equal to 1 means that all navigation paths are shortest possible. The plotted points are mean values, while the error bars show the minimum and maximum values obtained for the NNG over 10 random sprinklings of nodes for a given value of $\gamma$. Panel (b) shows the success ratio as a function of the percentage of edges randomly deleted from the network. The smaller the $\gamma$, the more robust the navigability concerning this network damage.




(a) (b)

Figure 5.11: NNG equilibrium networks share basic structural properties with real networks. The real networks considered are the Internet, metabolic reactions, and the English word network, see Section 5.6. Panel (a) and (b) shows the degree distribution and the average clustering coefficient of nodes of a given degree in the real and NNG networks. The dashed black lines are the power laws with exponents $-2$ and $-1$. The clustering coefficient of a node of degree $k$ is the number of triangular subgraphs containing the node, divided by the maximum possible such number, which is $k(k-1)/2$. In the NNG network, the disk radius is $R = 21.2$ and $\alpha = 0.5$. There are no other parameters.

| Network | Internet | Metabolic | Word | NNG |
|---|---|---|---|---|
| Nodes | 23748 | 602 | 4065 | 5000 |
| Edges | 58414 | 2498 | 38631 | 7955 |
| Avg. deg. | 4.92 | 8.29 | 19.01 | 3.18 |
| Avg. clust. | 0.61 | 0.55 | 0.45 | 0.60 |
| Avg. dist. | 3.52 | 3.22 | 2.43 | 3.89 |
| Diam. | 10 | 6 | 6 | 10 |

Table 5.2: Comparison of basic structural properties of real and NNG networks. The average distance and diameter are the average and maximum hop lengths of the shortest paths in the network. The average degree in the NNG-simulated network is lower than in the real networks because the NNG generates navigable networks with minimum numbers of edges. In the NNG network, the disk radius is $R = 21.2$ and $\alpha = 0.5$. There are no other parameters.

## 5.3 Network Navigation Game versus real networks.

Figure 5.12 and Table 5.3 show the results of our analysis applied to real networks. Panels (a), (b), and (c) on Figure 5.12 visualize the Internet, metabolic, and word networks mapped to the hyperbolic plane as described in the Section 5.6. The hyperbolic coordinates of nodes are then supplied to the minimum set cover algorithm that finds a Nash equilibrium of the NNG for each network. Panels (d) and (e) do the same for the US airport network and for the human brain, except that in the brain, the physical coordinates of nodes are used. The grey edges are present in the real networks but not in the NNG networks. These edges may exist in real networks for different purposes other than navigation so that the NNG can say nothing about them. The false-positive turquoise edges are present in the NNG networks but not in the real networks. The true positive magenta edges are present in both networks. Panels (f) and (g) show the NNG equilibrium network based on the physical (geographic, versus hyperbolic) coordinates of US airports and the NNG network for the Hungarian road network. The NNG networks have the same sets of nodes as the corresponding real networks, but the sets of edges are different. For visualization purposes, the grey edges are suppressed in the human brain and Hungarian road networks.

The detailed statistics of edges are in Table 5.3. We cannot expect real networks to be identical to NNG networks because the latter are minimum-cost maximum-navigation idealizations, while each real network performs many other functions different from navigation. In particular, since real networks must be error-tolerant and robust with respect to different types of network damage, we expect the number of edges in real networks to be noticeably larger then in their minimalistic NNG counterparts—something we indeed observe in Table 5.3. Yet if navigation efficiency does matter for real networks, then we should expect a majority of edges present in these NNG idealizations to be also present in the corresponding real networks. Table 5.3 confirms these expectations. The NNG precision in predicting links in real networks, defined as the ratio of NNG true positive links to the total number of NNG links, exceeds 80% for most networks, while the precision in predicting frame links, crucial for navigation, exceeds 90% for some networks. In what follows we juxtapose

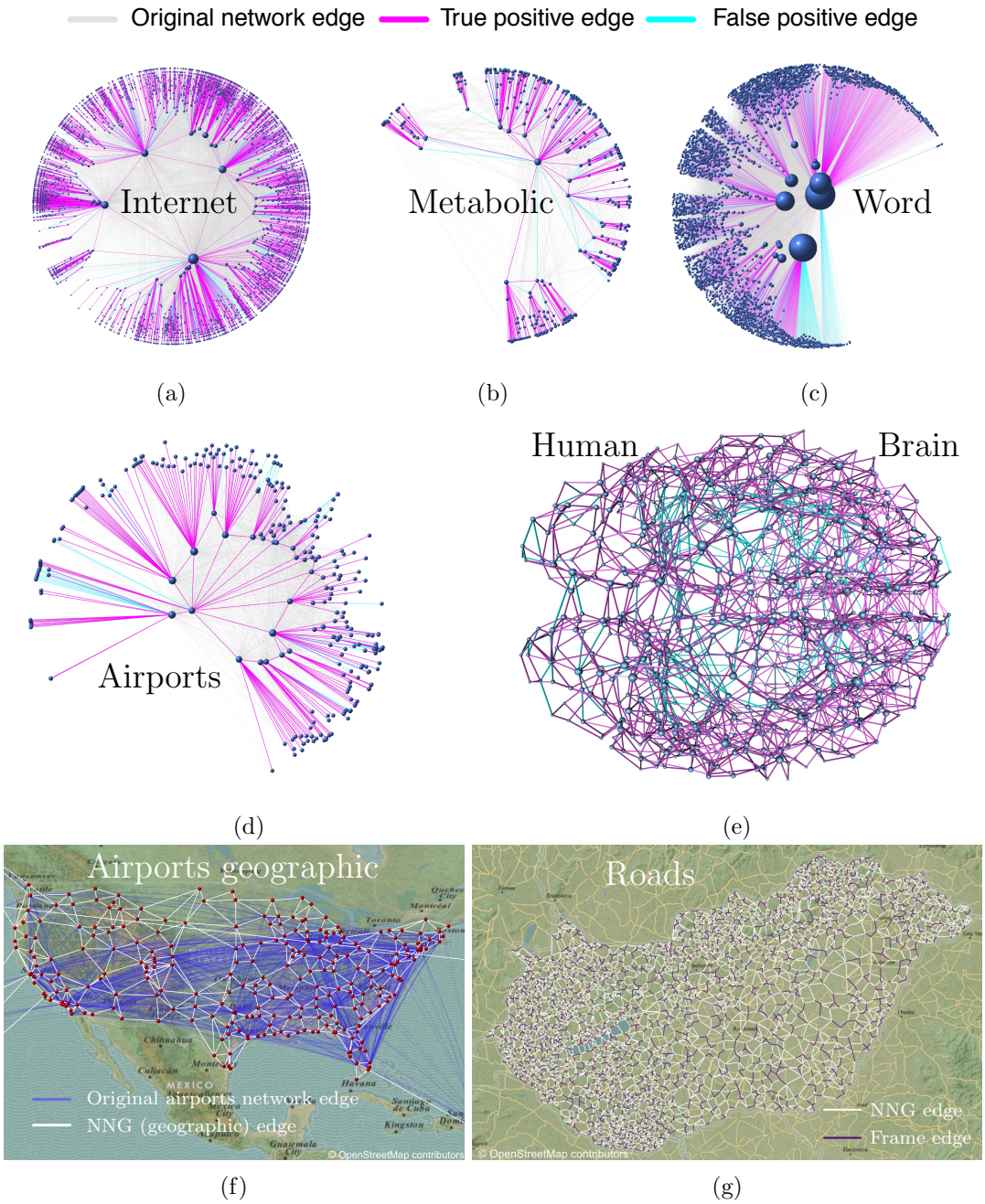



Figure 5.12: Network Navigation Game (NNG) predicts links in real networks.

| | Inter. H | Metab. H | Word H | Roads E | Airp. S | Airp. H | Brain E |
|---|---|---|---|---|---|---|---|
| Nodes | 4919 | 602 | 4065 | 3136 | 283 | 283 | 998 |
| Real edges ($\|R\|$) | 28361 | 2498 | 38631 | - | 1973 | 1973 | 17865 |
| NNG edges ($\|M\|$) | 5490 | 743 | 4634 | 9808 | 643 | 328 | 2591 |
| True positives ($\|T\|$) | 4556 | 643 | 3311 | 8776 | 65 | 277 | 2306 |
| False positives ($\|F\|$) | 934 | 100 | 1323 | 1032 | 578 | 51 | 285 |
| **Precision ($\|T\|/\|M\|$)** | 83% | 87% | 71.5% | 89.48% | 10.1% | 84% | 89% |
| Frame edges ($\|M_F\|$) | 3680 | 415 | 3304 | 3105 | 199 | 249 | 716 |
| Frame true positives ($\|T_F\|$) | 3243 | 378 | 2528 | 2931 | 15 | 216 | 677 |
| **Frame prec. ($\|T_F\|/\|M_F\|$)** | 88% | 91% | 77% | 94.40% | 7.5% | 87% | 94.6% |
| Navigation success ratio | 87% | 85% | 81% | - | 54% | 89% | 89% |

Table 5.3: The table quantifies the relevant edge statistics in Figure 5.12, showing the total number of edges in the real networks $|R|$, and in their NNG equilibrium networks $|M|$, the number of true positive (magenta edges in Figure 5.12) $|T|= |M \cap R|$, the number of false positive (turquoise edges in Figure 5.12) $|F|= |M \setminus R|$, and the true positive rate, or precision, defined as $|T|/|M|$. The precision statistics are also shown for the frame edges. Capital letters H,E,S after the network names refer to the embedding geometry: H:hyperbolic, E:Euclidean, S:spherical. The Euclidean coordinates in the brain are three-dimensional.

these numbers against the corresponding numbers in randomized null models, where they are exponentially small, upper bounded by 0.1%. Now, we provide probability estimates which represent the statistical significance of that the NNG equilibrium network links' containment by the real networks is very unlikely to occur by random chance, but rather is likely to be attributable to the specific characteristics of our embedding and NNG processes.

The NNG equilibrium network (graph) is a transformation of the real network under investigation by an embedding and a gaming (NNG) process. Although this transformation is completely deterministic, the statistical significance test can be performed in the following two ways: In the first approach the NNG equilibrium network is substituted by a completely random network with the same average degree $\bar{k}_{\mathrm{NNG}}$, that is $\frac{N}{2}\bar{k}_{\mathrm{NNG}}$ links are randomly chosen from the possible $\frac{N(N-1)}{2}$ number of links. The probability that $p$ fraction of these links (e.g. $p = 0.83$) are contained by the real network (having $\frac{N}{2}\bar{k}$ links) can be calculated as

$$\frac{\binom{N(N-1)/2-N/2\bar{k}}{(1-p)N/2\bar{k}_{\mathrm{NNG}}}\binom{N/2\bar{k}}{pN/2\bar{k}_{\mathrm{NNG}}}}{\binom{N(N-1)/2}{N/2\bar{k}_{\mathrm{NNG}}}} \tag{5.62}$$

which is in the order of $O(e^{-N})$. Because this probability is extremely small for reasonable $N$, our result is very unlikely to occur also along with fully random networks with fixing only the number of edges. For example, taking the values on the Internet AS-level topology embedding ($N = 4919, \frac{N}{2}\bar{k} = 28361, \frac{N}{2}\bar{k}_{\mathrm{NNG}} = 5490, p = 0.83$) the probability above is $5.62 \times 10^{-11068}$.

More refined randomization of the NNG equilibrium network is to substitute only the embedding process by fully random generation of H2 coordinates (with such coordinate distribution similar to the one resulted by the embedding process) and then apply the gaming process (as if the embedding was wrong and had no concern to the original real network). In this way, the resulted random NNG network preserves not only the average degree but the degree distribution and the clustering coefficient of the original NNG equilibrium network. Let $X$ be a random variable denoting the

number of links from the randomized NNG equilibrium network contained by the original real network. Inevitably, $X$ is a non-negative random variable bounded also from above by $P := \frac{N}{2}\bar{k}_{\mathrm{NNG}}$. Although the exact distribution of $X$ cannot be calculated due to the dependent link establishment of the gaming process, the expected value of $X$ (which is insensitive to link dependence) is

$$E(X) = \frac{N}{2}\bar{k}\frac{\frac{N}{2}\bar{k}_{\mathrm{NNG}}}{\frac{N(N-1)}{2}} \approx \frac{1}{2}\bar{k}_{\mathrm{NNG}}\,\bar{k}. \tag{5.63}$$

Based on this average value, a conservative upper bound can also be given on the probability that the level of this link containment exceeds a certain threshold $0 < C < P$. Applying Hoeffding's inequality [83] we can state that

$$P(X > C) \leq \left(\frac{E(X)}{C}\right)^{\frac{C}{P}} \left(\frac{P - E(X)}{P - C}\right)^{1-\frac{C}{P}} \tag{5.64}$$

This upper bound is far below 0.05 for several reasonable $\bar{k}$ and $N$. For example, the probability that more than 83 percent of the randomized NNG equilibrium network links ($C$ =4556 of the total 5490 edges) coincide real Internet edges (among the total 28 361) is upper bounded by 0.00136044. The complement of the upper bound of the probability above (1-upper bound) can also be considered as a weight of our statement (in the example above 0.99864).

We also note that since the real networks have many more links than NNG networks, their navigability may not suffer much from missing a small percentage of NNG links, as confirmed by the success ratio results in the same figure.

Of particular interest to us here are networks that are explicitly embedded in the physical space. In these cases, we may not need to embed the network, but use the physical coordinates of its nodes instead to construct the NNG equilibria. We consider three examples: the Hungarian road network, the airport network of the United States, and a structural network of the human brain. In the first network, the nodes are the cities, towns, and villages of Hungary, while in the second network, the nodes are US airports. Two nodes are linked if they are connected by a direct road or flight. In the brain network, the nodes are small regions of an average size of $1.5\mathrm{cm}^2$ both hemispheres of the cerebral cortex entirely, and two regions are connected if a structural connection between them is detected in diffusion spectrum imaging. We expect the NNG to be particularly accurate in predicting links in these networks using the physical—instead of hyperbolic—coordinates of nodes. We note that these physical coordinates are Euclidean in all three cases. The embedding space is two-dimensional Euclidean and spherical space in the road and airport cases, and it is three-dimensional Euclidean space in the brain case. Our method to construct an NNG equilibrium applies without change to any set of points in any geometric space. For example, we show analytic results on the structure of NNG equilibrium networks in Euclidean spaces.

**Results for the Euclidean space**

We analyze the degree distribution in NNG equilibrium networks constructed on sets of points sprinkled uniformly at random over Euclidean disks. We show that the expected degree of a node located in the disk center is around 1, while the expected degree of a node at the disk boundary is around 1/2. Because of this lack

of variability of node degrees, the degree distribution in the Euclidean case cannot have any fat tails.

According to (5.6) the expected degree of node $u$ is

$$\delta \int_0^R \int_0^{2\pi} e^{-\delta T_{uv}} \mathrm{d}\phi r_v \mathrm{d}r_v \ , \tag{5.65}$$

where $\delta = N/T_R = \frac{N}{R^2\pi}$. To give an upper bound we will give a lower bound for $T_{uv}$. If $u$ is the centre of the disk, then $T_{uv}$ is the area of the intersection of the disk and an circle around $v$ with radius $r_v$. If $r_v \leq R/2$, then this intersection is the circle itself around $v$, else the intersection contains a circle with radius $R/2$, hence

$$\begin{aligned} k(0) &\leq \delta \int_0^{R/2} \int_0^{2\pi} e^{-\delta r_v^2 \pi} \mathrm{d}\phi r_v \mathrm{d}r_v + \delta \int_{R/2}^R \int_0^{2\pi} e^{-\delta (R/2)^2 \pi} \mathrm{d}\phi r_v \mathrm{d}r_v \\ &\leq 1 - e^{-\frac{1}{4}\delta R^2 \pi} + \frac{3}{4}\delta R^2 \pi e^{-\frac{1}{4}\delta R^2 \pi} \leq 1 + 3\frac{N}{4} e^{-N/4} \leq 1 + \frac{3}{e}. \end{aligned} \tag{5.66}$$

Moreover, if $N \geq 6$ then $k(0) \leq 1.05$

To give a lower bound on the expected degree we will count with the whole circle around $v$ instead of the intersection:

$$k(0) \geq \delta \int_0^R \int_0^{2\pi} e^{-\delta r_v^2 \pi} \mathrm{d}\phi r_v \mathrm{d}r_v = \delta 2\pi \int_0^R e^{-\delta r_v^2 \pi} r_v \mathrm{d}r_v = 1 - e^{-\delta R^2 \pi} = 1 - e^{-N}. \tag{5.67}$$

If $N \geq 6$, then $k(0) \geq 0.99$.

Similarly, for the expected degree of a node $u$ at the disk boundary

$$k(R) \geq \delta \int_0^R \int_0^{2\pi} e^{-\delta d^2 \pi} \mathrm{d}\phi r_v \mathrm{d}r_v, \tag{5.68}$$

where $d$ is the distance between $u$ and $v$, and according to the cosines law, $d^2 = R^2 + r_v^2 - 2Rr_v \cos\phi_v$. The inner integration is

$$\int_0^{2\pi} e^{-\delta\pi(R^2 + r_v^2 - 2r_u r_v \cos\phi_v)} \mathrm{d}\phi = 2\pi \mathrm{I}(0, 2\pi\delta r_v R) e^{-\delta\pi(R^2 + r_v^2)}, \tag{5.69}$$

where $\mathrm{I}(0, x)$ is the BesselI function. Unfortunately the BesselI cannot be integrated, but we can use that $\mathrm{I}(0, x) \sim e^x/\sqrt{2\pi x}$. Hence

$$\begin{aligned} k(R) &\geq \int_0^R \frac{2\pi\delta}{\sqrt{4\pi^2 \delta R r_v}} e^{2\pi\delta R r_v - \delta\pi(R^2 - r_v^2)} r_v \mathrm{d}r_v = \int_0^R \frac{r_v N}{\sqrt{R^3\pi}} e^{-\pi(R - r_v)^2 N/R^2} \mathrm{d}r_v \\ &\geq \frac{2\sqrt{N}}{3\sqrt{\pi}} \mathrm{HypergeometricPFQ}\left(\left\{\frac{1}{2}, 1\right\}, \left\{\frac{5}{4}, \frac{7}{4}\right\}, -N\right) \xrightarrow[N\to\infty]{} \frac{1}{2} \end{aligned} \tag{5.70}$$

On the left panel of Figure 5.13 the simulation results support the analytical findings that in the Euclidean case the expected degree nodes as a function of their radial coordinates have very low variability in the NNG equilibrium networks and their frame topologies. As a consequence of this low variability the degree distributions do not have any fat tails or power laws and decay fast with the node degree,

the right panel of Figure 5.13. Clustering is still relatively strong however: in the synthetic Euclidean NNG network it is 0.19, in the road NNG network it is 0.22, while in the brain network and its NNG, the clustering values are 0.46 and 0.21, respectively.

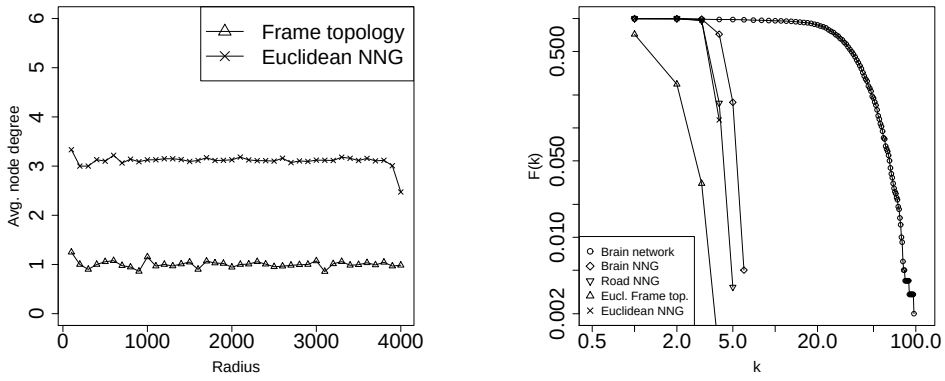
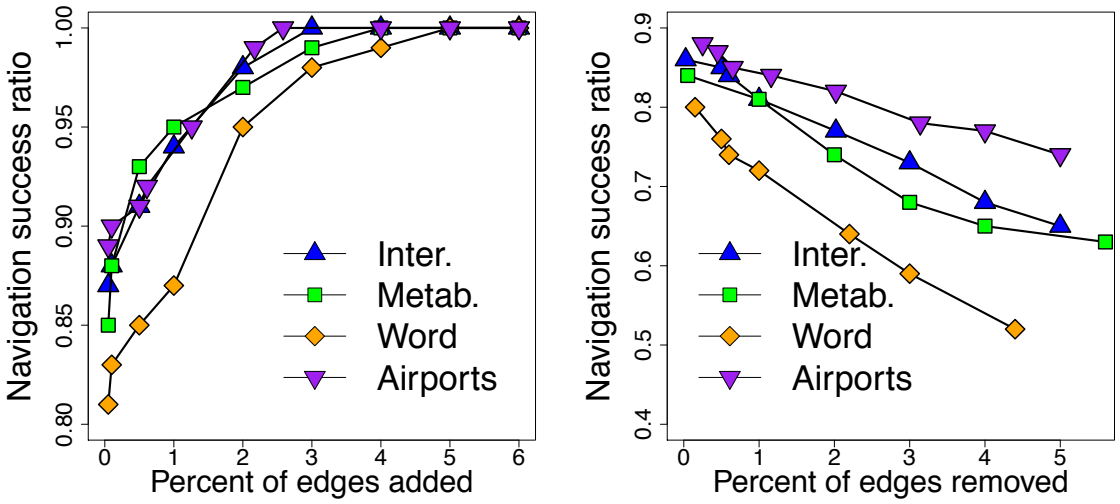



Figure 5.13: The average degree of nodes as a function of their radial coordinates on a Euclidean disk (left), and the cumulative distribution function of node degrees in the corresponding NNG equilibrium, its frame topology, the Hungarian road network, the brain network, and its NNG equilibrium (right).

We apply our method to find the NNG equilibrium networks using the physical coordinates of nodes in these three real networks, and then compare them to their NNG equilibria also in Figure 5.12 and Table 5.3. We observe that in the brain and road networks, the NNG link prediction accuracy is particularly high, reaching 89% for all the links and 94-95% for the frame links. For the brain, this result implies that the spatial organization of the brain is nearly optimal for information transfer, in agreement with previous results [164, 103, 82, 72]. In the Hungarian road network, nearly all frame links, crucial for efficient navigation using geography, are present. Practically this means that Hungarians have the luxury to go on a road trip without a map since all the major roads required by geographic navigation are there, albeit the condition of some of those roads is not as luxurious. Simply put, there are roads where people with a compass may think they should be.

For the US airport network, however, the geographic results are poor. These poor results may be unexpected at first, but they have a simple explanation in that the geometry of the airport network is not really Euclidean, as the geometry of the nearly planar road network, but hyperbolic. Indeed, efficient paths in the airport network optimize not so much the geographic distance traveled, but the number




(a) (b)

Figure 5.14: NNG equilibria of real networks helps to improve or degrade their navigability. The edges from the NNG equilibria of the considered real networks are first sorted in the decreasing order of betweenness centrality, and then either added to the real network if not already there (panel (a)), or removed from the network if present (panel (b)). The $x$-axis shows the percentage of added or removed edges compared to the number of edges in the original real network. The navigation success ratio is computed as the number of node pairs between which geometric routing is successful, divided by the number of all node pairs.

of connecting flights. As a consequence, most paths go via hubs. As opposed to the road network, where the number of roads meeting at an intersection does not vary that much from one intersection to another, the presence of hubs in the airport network makes the network heterogeneous, i.e., node degrees vary widely. This heterogeneity effectively creates an additional dimension (the "popularity." dimension in [141]). That is, in addition to their geographic location, airports also have another important characteristic—the size or degree. This extra dimension makes the network hyperbolic [101]. The NNG results for the hyperbolic map of the airport network in Figure 5.12 are as good as for the other networks.

## 5.4 How to cure or injure a network efficiently.

The knowledge of the NNG equilibrium of a given real network makes it possible to efficiently identify links that are most critical for navigation in the network. Since NNG equilibrium networks are maximally navigable networks composed of the smallest number of links, we expect that if we alter a real network by either adding or removing a relatively small number of links belonging to the NNG equilibrium of the network, then such network modifications may significantly affect network navigability.

Figure 5.14 supports these expectations. In the figure, we take the considered real networks, and add to them certain numbers of links that are present in the NNG equilibria of the real networks, but not present in the networks themselves. About

1-2% of added edges, compared to the original numbers of edges in the networks, increase network navigability significantly, while the addition of 2-5% of edges make all the networks 100%-navigable. Similarly, the targeted removal of a small portion (1-5%) of edges belonging both to the NNG equilibria of the networks, and the network themselves, degrades network navigability by 10-30%.

## 5.5 Discussion

We emphasize that the considered Nash equilibrium networks are minimalistic idealizations, concerned only with maximizing the efficiency of the navigation function at minimal cost (number of links). Reality differs from this ideal in many ways. First, real networks must be robust concerning noise and random failures. This robustness requirement explains why the considered real networks have strictly more links that their Nash equilibria. Maximum navigability can be achieved not only at the minimal cost but also at a higher cost. Second, transport processes in real networks are also noisy and can follow not only steepest descent path (greedy navigation) but also any downstream paths, still achieving 100% reachability. Yet the noisier the transport process, the less likely it stays to the shortest path, leading to higher stretch and longer travel times, thus degrading navigation efficiency in terms of these parameters. Third, navigability does not always have to be maximized as many specific networks perform many specific functions other than navigation. Our game-theoretic approach can be extended to accommodate some of these functions, such as error tolerance or policy compliance [161], but not all possible functions of different real networks can be formalized within this game-theoretic framework. Some networks are centrally designed to optimize a particular function globally [104]. Game theory is not needed to formalize such global optimization strategies. It is more suited for self-organized networks, in which each node behaves selfishly according to its incentive, independent of other nodes. In other words, Nash equilibrium networks are structural manifestations of local incentives of nodes for efficient transport or communication, in contrast with existing generative or optimization models of complex networks [50, 126]. Finally, all real networks are dynamic and growing, while Nash equilibria correspond to static network configurations. However, it has been recently shown [100] that in case of random geometric graphs—to which the considered Nash equilibrium networks effectively belong according to the following results —one can map an equilibrium network model to an identical growing one. The Heaviside step function with the step at

$$R' = 2 \ln \frac{\bar{k}}{8\delta} \tag{5.71}$$

is a good approximation to the effective connection probability in Eq. (5.15) for $\delta \in [10^{-6}, 10^{-3}]$ and $R = [12, 18]$. With this step-function approximation, node $u$ connects to $v$ iff $d_{uv} \leq R'$. Therefore the expected degree of $u$ is the expected number of points lying within the intersection of the $R-$disk and the $u$-centred disk of radius $R'$.

To see that this step function is indeed a good approximation to the effective connection probability in the NNG equilibrium, recall that the area of the two disks

above can be approximated as

$$T_{R',R} = 4e^{\frac{R'}{2}} e^{\frac{R-r_u}{2}}. \tag{5.72}$$

From these one can obtain

$$k_{\text{out}}(r_u) \approx N \frac{T_{R',R}}{T_{R-\text{disk}}} = N \frac{4e^{\frac{R'}{2}} e^{\frac{R-r_u}{2}}}{\pi e^R} \ . \tag{5.73}$$

If $R'$ from (5.71) is substituted into the formula above we get back the expected out-degree in (5.20). In particular, if $R' = R$ (as in [102]), then

$$k_{\text{out}}(r_u) = \frac{4}{\pi} N e^{-\frac{r_u}{2}} \tag{5.74}$$

and

$$\bar{k} = \frac{8}{\pi} N e^{-\frac{R}{2}}, \tag{5.75}$$

which coincides with Eqs. (12,13) in [102].

Notwithstanding these limitations, we have shown that ideal networks designed to be maximally navigable at minimal cost, share basic structural properties with real networks. Compared to existing works on navigation-optimal distributions of shortcut edges in Euclidean grids [92, 108, 151, 107] which do not yield realistic network topologies, this result is quite unexpected because there is absolutely nothing in the definition of these ideal networks that would enforce or even welcome a formation of any particular network structure. The networks are defined purely in terms of navigation optimality. The surprising finding that the structure of these ideal networks is similar to the structure of real networks should not be misinterpreted as if these idealizations are generative models for real navigable networks. Instead the former are skeletons or subgraphs of the latter. Since these skeletons consist of the minimum number edges required for 100% navigability, there is no even a parameter to control the most basic structural network property—the average degree, which is always controllable in generative models. On the contrary, as follows from Eq. (5.23), the average degree in these skeletons is uncontrollable and lies between 1 and 4.

We find that if network geometry is hyperbolic, then our navigation skeletons have power-law degree distributions and strong clustering. The values of power-law exponent $\gamma$ close to 2, observed in many real networks [131, 23], appear as the best possible choice. In this case, not only reachability is 100%, but also the network cost and stretch are minimized and navigability robustness is maximized, compared to other values of $\gamma$ in Figure 5.10.

These results apply to sets of points in hyperbolic space, but the navigation skeleton construction itself is by no means limited to these hyperbolic settings. It is very general and applies to any set of points in any geometric space, as illustrated by the brain and road networks where we have used the Euclidean $2d$ and $3d$ physical coordinates of nodes to construct the navigation skeleton of the network. Our finding that the brain contains almost fully its navigation skeleton appears as a mathematically clear and conclusive evidence that the spatial organization of the

brain is nearly optimal for communication and information transfer, corroborating existing work on the subject s [164, 103, 82, 72].

We note that the connection between the structure and function of networks are often studied in the logically reverse direction: structure→function. That is, first some data about the structure of real networks is obtained, and then questions concerning how optimal this structure is with respect to a given network function are investigated. This logic does provide some evidence that the network might have evolved optimizing this function, but this evidence is quite indirect and unreliable compared to the direct demonstration that functionally optimal networks have the structure observed in reality: function→structure. Common sense suggests that this causal direction must reflect reality more adequately since networks, either designed or naturally evolving, do not have a completely random structure but the structure (effectively) optimizing some functions. Yet studying networks in this direction is much more challenging primarily because of difficulties in formalizing the constraints that a given function imposes, and deriving the resulting optimal network structure. Here, with the help of game theory, we have done so for the navigation function that many real networks (implicitly) perform.

As one would logically expect, the function→structure approach provides a deeper insight into specific details of the network's structural organization that is critical for its functional efficiency. We have confirmed this expectation by demonstrating that our approach can identify links in real networks that are most critical for navigation. A targeted attack on these critical links degrades navigability rapidly, while if a real network is not 100%-navigable, our approach finds the minimal number of not-yet-existing links whose addition to the network boosts up its navigability to 100%. Therefore our approach can be used to identify real network links that should be protected most in a critical network infrastructure. On the other hand, this approach can also, help network designers to prioritize possible link placement options, i.e., pairs of not directly connected nodes, that, if connected, would maximize navigability improvement.

Finally, all the real networks considered here are expected to be navigable. Indeed, the primary functions of the Internet, brain, metabolic, or airport and road networks are to transport information, energy, or people. Semantic and syntactic navigability of word networks is an established fact in cognitive science [63, 41, 14]. However one cannot expect all real networks to be highly navigable as navigation is not an important function of every network in the world. One example is a technosocial web of trust, in which nodes are public keys of users of a distributed cryptosystem, linked by users' certifications of key-user bindings. There is no reason why this network should be navigable. In agreement with this observation, we then find that this network does not contain a large percentage of edges from its NNG equilibrium, suggesting that the introduced methodology can also be used as a litmus test to investigate if navigation is an important function of a given real network, and if so, then to what degree. One cannot expect every real network to be highly navigable because navigation is not an important function of every real network. Here we consider one example, the Pretty-Good-Privacy (PGP) web of trust network, specifically the December 2006 snapshot and its hyperbolic coordinates from [141]. These data are then processed exactly as for all the other networks. However, as expected, the navigation success ratio and precision metrics reported

for this network in Table 5.4 are substantially lower than for the navigable networks.

| | PGP |
|---|---|
| Nodes | 4899 |
| Real edges ($\|R\|$) | 67650 |
| NNG edges ($\|M\|$) | 29311 |
| True positives ($\|T\|$) | 6945 |
| False positives ($\|F\|$) | 22366 |
| **Precision ($\|T\|/\|M\|$)** | 24% |
| Navigation success ratio | 36% |

Table 5.4: The table quantifies the relevant edge statistics showing the total number of edges in the core of the PGP network $|R|$, and in its NNG equilibrium network $|M|$, the number of true positive edges $|T| = |M \cap R|$, the number of false positive green edges $|F| = |M \setminus R|$, and the true positive rate, or precision, defined as $|T|/|M|$.

## 5.6 Technical details

**The real network data.** The Internet dataset representing the global Internet structure at the Autonomous System (AS) level is from [27]. The metabolic network is the post-processed network of metabolic reactions in *E. coli* from [141], Snapshot $S_1$ there. The post-processing details can be found in [141]. The word network is the largest connected component of the network of adjacent words in Charles Darwin's "The Origin of Species" from [120]. The airport network was downloaded from the Bureau of Transportation Statistics `http://transtats.bts.gov/` on November 5, 2011. The structural human brain network and physical coordinates of nodes (regions of interest (ROIs)) in it are the diffusion spectrum imaging (DSI) data from [80].

**The hyperbolic maps of real networks.** The hyperbolic coordinates of ASes and metabolites are from [27] and [141]. The hyperbolic coordinates of words and airports are inferred using the *HyperMap* algorithm [140]. This algorithm is deterministic and is based on the growing network model in [141] used to show that the latent geometry of scale-free strongly clustered real networks is hyperbolic. Given an adjacency matrix of a real network, the algorithm infers the hyperbolic coordinates of its nodes by replaying its growth as the model in [141] prescribes. Accurately, the nodes are first sorted in the order of decreasing degrees, and then, starting with the highest-degree node, nodes and their edges are added, one node at a time, to a growing network. The probability, or the likelihood, with which model [141] generates this growing network, depends on the node coordinates. The HyperMap algorithm sets the coordinate of each added node to the coordinate corresponding to the global maximum of this probability.

**The Nash equilibrium networks of NNGs.** The hyperbolic or physical, in the airport and brain cases, coordinates are then supplied to the GNU Linear Programming Kit (GLPK) `http://www.gnu.org/software/glpk/` used to find a solution to the corresponding minimum set cover problem. To yield acceptable

running times of the solver, the Internet and word networks are reduced in size by extracting their high-degree cores of about 4500 nodes. The Hungarian road data is processed slightly differently. First, the cities in Hungary are mapped to their geographic coordinates using the database in `http://www.kemitenpet.hu/letoltes/tables.helyseg_hu.xls`. Then these coordinates are used in the GLPK to find the NNG equilibrium. Each edge in this equilibrium network is then checked for existence in the real road network. To check that, the GoogleMaps API `https://pypi.python.org/pypi/googlemaps/` is used to find the shortest path between the two cities connected by the edge. The edge is defined to also exist in the real road network if this shortest path does not go via any other city.

# Chapter 6

# Hierarchical systems

We have seen so far that greedy navigation, supported by the hidden metric space of the network, can account for the excellent navigability of networks. Although the framework of greedy navigation is very compelling, the embedding of real networked systems into metric spaces ensuring reliable navigation can be very cumbersome and non-intuitive in many cases (see [25]). In such cases, the function→structure approach clearly inherits the non-trivialities of greedy navigation and metric spaces.

In this chapter, we show that the characterization of navigation paths used in networks can be achieved to a sufficient extent. This enables the function→structure analysis without assuming the mechanism of greedy navigation. Our approach here focuses on the high-level structure of the paths used in the network. There are numerous examples that real networks exhibit a hierarchical structure. Organizational (e.g., military) networks, for example, are well-known to have a clearly defined hierarchy. The Internet is another example in which the connections between internet domains are hierarchical, pointing from customer to provider. We show that these underlying hierarchies have a significant impact on the operative paths in the network. At this point, we turn back to our networks and paths investigated in Chapter 3, the Internet AS topology, the air transportation network, the word morph network, and the human brain. Recall that for these networks, we have collected large datasets about both the structure of the networks and the empirical paths. A deeper analysis of these empirical paths uncovers two additional features (on top of stretch introduced in Chapter 3) in connection with the underlying hierarchy.

One such feature our measurements support is "conform hierarchy." (CH), meaning that the used paths follow the topological hierarchy of the network. For showing this, we have computed the closeness centrality of the nodes comprising the empirical paths indicating which (inner or outer) parts of the network the information flows through. The closeness centrality of the node is computed as: $C(x) = \frac{N}{\sum_y d(y,x)}$, where $d(y, x)$ is the distance between vertices $x$ and $y$, while $N$ refers to the number of nodes in the network. We found that most of the empirical paths do not contain a large-small-large pattern forming a " valley" anywhere in their closeness centrality sequence. This informally means that higher-level nodes do not prefer the exchange of information through their subordinates, even if there are short paths through them. On a CH path, the closeness centrality increases monotonically at first up to a point (upstream), then starts to decrease (downstream) until it reaches the destination, or it is just going upstream or downstream all the way. Fig. 6.1 illus-

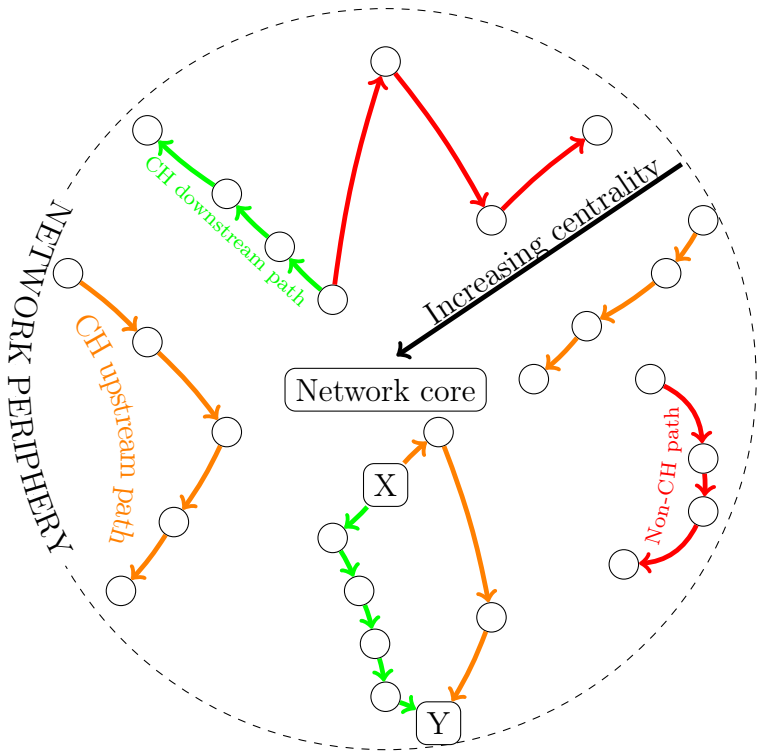



Figure 6.1: Illustration of paths with regard to the internal logic of the network. A path is CH if it does not contain a large-small-large pattern forming a "valley" anywhere in its centrality sequence (green and orange paths). Red paths show examples of non-CH paths. An upstream path contains at least one step upwards in the hierarchy of the network (orange paths), while in downstream paths, the centrality decreases all the way (green paths).

trates this graphically. One could argue that maybe short paths on real networks have this property as a default, but Fig. 6.2**a-d** verify that this is not the case. For comparison, we picked random paths between the source-destination pairs of our empirical paths with the same stretch distribution and plotted the results for that case too. One can see that, while the path length distribution is the same for the two datasets, a much larger fraction of stretch-equivalent random paths violate the CH feature.

There can be subtle differences between CH paths of similar length. For example, a path can contain upstream than downstream steps or downstream steps only. Recall that an upstream step goes towards the core, while a downstream step goes towards the periphery of the network. Is there a preference among these? For answering this, we plotted the Cumulative Distribution Function (CDF) of CH paths with respect to the number of upstream steps preceding the downstream phase (Figure 6.3**a-d**). For comparison, we have also plotted the results of a random policy that picks randomly from the possible CH paths of the given length. The plots confirm that the empirical paths contain much less upstream steps, which means that these paths try to avoid stepping towards the core. This finding adds "prefer downstream" as a third identifiable path selection feature (see Fig. 6.1 for an illustration). We note that such behavior is easy to interpret on the Internet, since

stepping towards the core of the network implies paying for a transit provider for carrying the traffic while going downstream comes for (almost) free. However, at this time, it is not clear what causes the same behavior in the other networks.

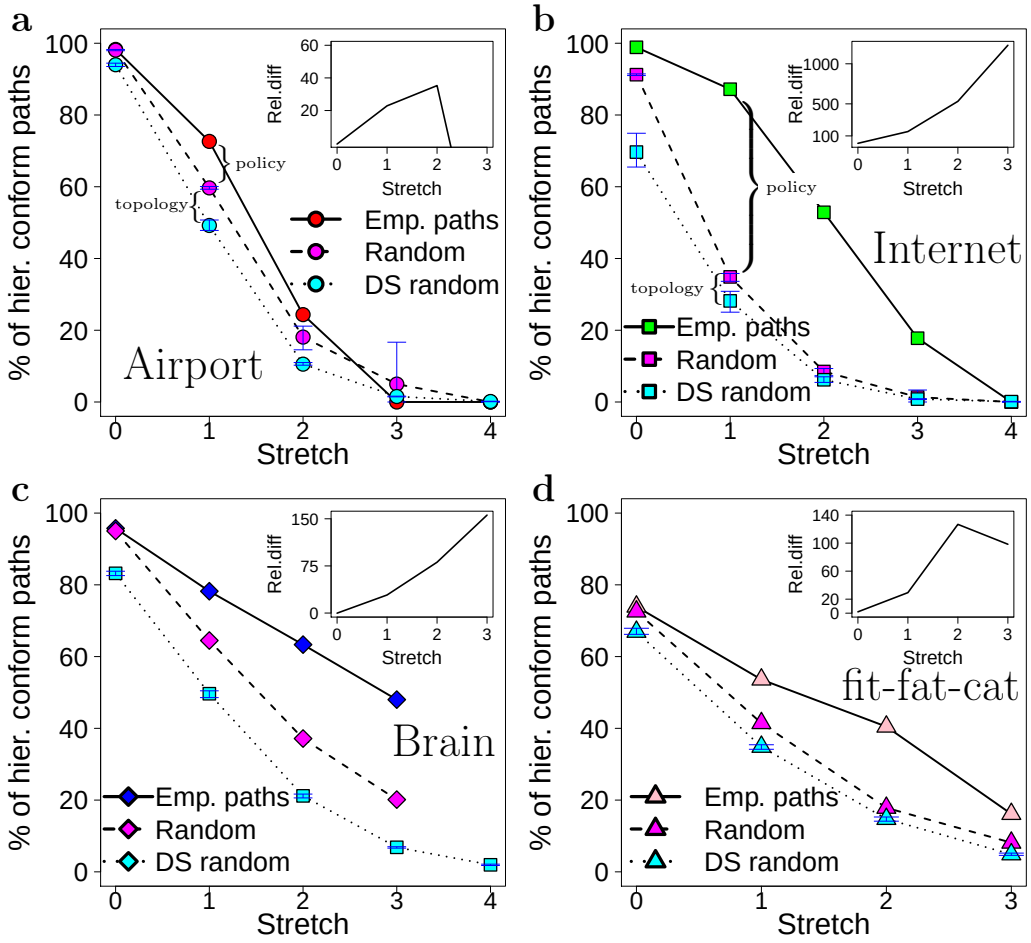



Figure 6.2: Identified path selection features confirmed by our measurement data. Panels **a-d** show the hierarchical conformity of the empirically-determined paths against stretch. The inset of the plots shows the relative difference between the number of CH paths in the empirical and random paths. In the case of small networks, there are 15-85% more CH paths in the empirical traces, but in the case of the large AS level Internet, this goes up to 100-500%. The cyan colored data in the plots show the number of CH paths in a randomized version of our networks generated with the degree sequence (DS) algorithm, which produces exactly the same degrees for the nodes, but the edges are completely randomized. The plots confirm that the topological peculiarities of real networks increase the number of CH paths between endpoints with respect to the DS networks (see the explanation brackets between the cyan and magenta-colored dots of panel **a** and **b**). However, we argue that the effect of the CH feature is at least that important or even more fundamental (e.g., in case of the Internet).

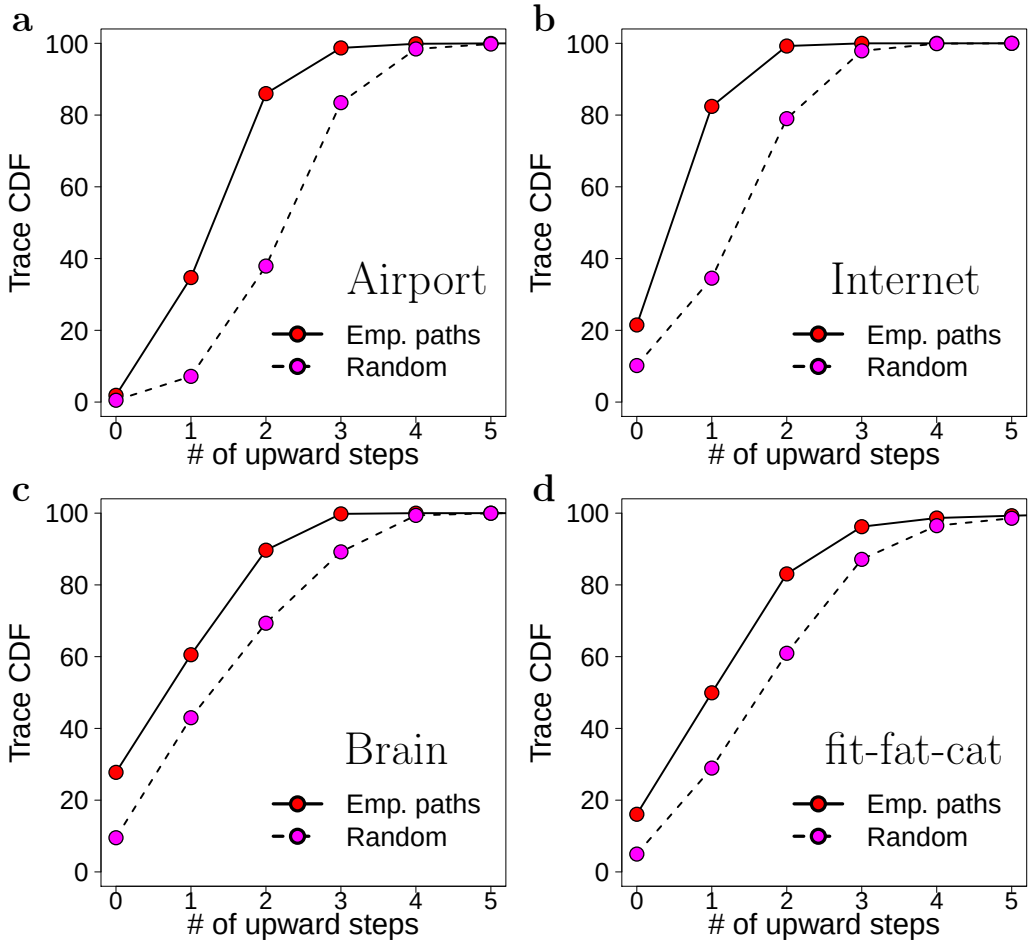



Figure 6.3: Panels **a-d** show the cumulative distribution of upstream steps in the traces of our datasets. The empirical paths tend to avoid stepping towards the core, which is reflected by the much lower number of upstream steps (in comparison with the randomly selected CH paths of the same length) before entering the downstream phase.

## 6.1 Function-structure analysis of the Internet

The Internet with its intrinsic customer-provider hierarchy is the pathological example of networks clearly built under hierarchical relationships. Our current understanding of the domain-level topology of the Internet is based on measurements and generative models which set up rules describing the behavior (node and edge dynamics) of the individual ASes and generalize the consequences of these individual actions for the complete AS ecosystem. Here we apply the function→structure approach on the Internet, based on our observation of hierarchies influencing path selection in networks and the Internet's policy routing ecosystem. We show that such a function→structure approach can give complementary insights into the topological properties of the AS network. In contrast with generative models reflecting high-level statistics (e.g., degree distribution, clustering, diameter), our reasoning can identify omnipresent subgraphs and peering likelihood.

One cannot overestimate the value of knowing more about the topology of the In-

ternet. The last decades have supplied us with thousands of stories where topology-related information about the Internet was directly transformed into more efficient architectures and services, or more appropriate business decisions. The most specific example is clearly Content Delivery Networks (CDN) [143], where global topological peculiarities are highly exploited, e.g. in surrogate and cache placement strategies or request routing mechanisms [143] but CDN is just a narrow segment of the whole spectrum. The placement of data centers [74], peer-to-peer networks [39, 112], traffic engineering [9], business based AS peering strategies [44], just to mention a few, can clearly benefit from Internet topology related knowledge. The investigation of the AS topology is also a popular topic [26, 25, 6, 38, 173, 115, 149] in the network science community, which consolidates researchers from diverse or multidisciplinary research areas. One reason behind this popularity is that compared to other complex networks, active and passive measurements can be executed on the Internet topology, thus we can create Internet "screenshots" easily.

With this non-comprehensive list of consumers in mind, it comes at no surprise that many researchers, even from diverse or multidisciplinary research communities have contributed to our current understanding of the Internet's Autonomous Systems (AS) level topology. The only way to obtain ground-truth data about the AS topology is via active or passive measurements. Today we have historical and contemporary measurement data collected continuously and made publicly available according to various approaches (e.g., using BGP info [33, 167], traceroute measurements [155], IXP anatomy [3]). Meanwhile, the data stemming from these measurements are the exclusive source of direct information about the AS topology and thus can be treated as the ground truth, we can keep ourselves to, the way these measurement systems work is continuously reported to be imperfect and far from optimal [3]. Additionally, the collected data reveals only the current state of the network and cannot give usable predictions and clear characterization of the topology-forming processes lying in the background. Over the last four decades the Internet has evolved from a carefully engineered computer network, connecting universities and research institutes in the US, into a complex ecosystem on top of an overwhelming variety of stakeholders all over the world. The network science community emphasizes mostly the resemblance of the AS network to many real-world self-organizing networks, which is clearly the case but we argue that this network also has a second face as it apparently exhibits topological peculiarities stemming from technological underpinnings (e.g., the used networking technologies and protocol stacks). The underlying interdomain routing protocol guiding path selection on the Internet provides a good starting point for finding an analytically tractable set of path features for our function→structure analysis while providing non-trivial insights into the lineament of the AS network's technological face.

The interdomain routing policies of all the ASes are expressed through the well-defined framework of the Border Gateway Protocol (BGP) [146]. The main responsibility of BGP is to distribute the available forwarding paths between ASes and let them select their preferred paths according to their special interests. Table 6.1 recalls a simplified version of the usual steps of the path selection process in BGP from [71]. Here we highlight the vital *valley-free* criteria as a rule No. 0 since BGP path selection works over valley-free paths, which means that paths have to conform to the Internet's underlying customer-provider hierarchy. On top of the valley-free

Table 6.1: The simplified BGP best path selection process.

| # | Rule |
|---|---|
| 0. | Valley-free route |
| 1. | Highest local preference |
| 2. | Shortest AS path |
| 3. | Lowest origin type |
| 4. | Lowest MED |
| 5. | eBGP-learned over IBGP-learned |
| 6. | lowest IGP metric to the BGP next-hop |

feature, we also include the local preference rule, which formalizes the prefer downstream feature (see our measurements at the beginning of Chapter 6) adapted more rigorously to the context of the Internet.

### 6.1.1 The Internet's path selection policy

In the AS ecosystem, the business relationships between ASes can be quite diverse, still we can classify most AS-AS links into basically two major groups [86]: in a *customer-provider* relationship the customer AS pays the provider for forwarding its traffic, while in a *peering* relationship neighboring ASes voluntarily exchange traffic with each other in a settlement-free manner[1]. The valley-free policy manifests the simple economic principle that the flow of traffic must coincide with the flow of cash. In very short the policy dictates that AS $A$ can use a link to a neighboring AS $B$ to forward the traffic if and only if either the incoming traffic is from a customer, or $B$ is a customer of $A$. Putting it differently valley-free compliant paths comprise arbitrary (maybe zero) number of customer-provider links, zero or one peer link and again arbitrary provider-customer links strictly in this order (Fig. 6.4).

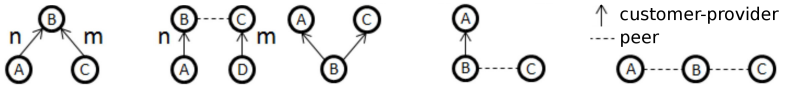



Figure 6.4: Illustration of valid (a) and invalid (b) valley-free path types. A valid path contains $n$ customer-provider, at most 1 peer and $m$ provider-customer link strictly in this order, where $n, m \in \mathbb{N}$. All the other types are invalid paths.

The local preference policy is applied on top of valley-free routes meaning that an AS can pick one from the available valley-free routes according to its local interest. Meanwhile, these local interests can exhibit great variety the minimalistic rule that customer and peer paths are favored over provider paths, is contained in basically every local preference setting within the ASes. This is in line with the nature of these routes as customer and peer paths are entirely free unlike provider paths in which the provider has to be compensated in some way for the carried transit traffic.

[1]We omit sibling and backup relationships for simplicity.

### 6.1.2 Formulation of the function-structure approach to the Internet

Similarly to the function→structure analysis of navigable networks in Chapter 5, in the followings we think about the ASes as rational but selfish players whose incentive is to communicate with each other using the valley-free and local preference policies. On top of these incentives we define the Hierarchical Network Game (HNG). Formally, let $\mathcal{P}$ be the set of players (identified as the ASes) with cardinality $N$. According to the valley-free rule an edge connecting two nodes $u, v$ can be of type either $\overline{uv}$ or $\overrightarrow{uv}$, where $\overline{uv}$ denotes a *peer* edge and $\overrightarrow{uv}$ denotes a *customer-provider* edge. The strategy space for node $u \in \mathcal{P}$ is a vector of the preferred edges to other nodes in the AS network, i.e., the set $S_u = \{(s_{uv})_{v \in \mathcal{P} \setminus \{u\}} : s_{uv} \in \{0, p, r\}\}$ where $|S_u| = 3^{N-1}$ and $p$, $r$ refer to $\overrightarrow{uv}$, $\overline{uv}$ edges, respectively. Easily, node $u$ seeks to contact node $v$ if $s_{uv} \in \{p, r\}$, otherwise $s_{uv} = 0$. We assume simultaneous announcement of the strategies between the nodes. Any state of the game is represented by an undirected graph $G(s) = (\mathcal{P}, E(s))$ generated by the strategies of the nodes $s$, where $E(s)$ is given by $E(s) = \{\vec{uv} | s_{uv} = p \wedge s_{vu} = 0\} \cup \{\overline{uv} | s_{uv} \in \{r\} \wedge s_{vu} \in \{r\}\}$. This settlement of the edges reflects the rational behavior of the ASes as they prefer to create peer edges over customer-provider edges.

The goal of the nodes is to minimize their costs which for a given node $u$ we define as:

$$C_u(s) = \underbrace{\frac{1}{N} \sum_{\forall v \neq u} d_{G(s)}(u, v)}_{\text{communication cost}} + \underbrace{\varphi_p u_p + \varphi_r u_r}_{\text{maintenance cost}}, \quad v \in \mathcal{P} \tag{6.1}$$

where

$$d_{G(s)}(u, v) = \begin{cases} 0 & \text{if exists a valley free path of which first edge is } \textit{peer} \text{ or } \textit{provider-customer} \\ 1 & \text{if exists at least one valley free path and the first edge of all of them is } \textit{customer-provider} \\ \infty & \text{if valley free path does not exist} \end{cases} \tag{6.2}$$

represents the price of communication between $u$ and $v$ over $G(s)$ in compliance with the policies defined, $\varphi_p$ and $\varphi_r$ are fix maintenance costs of the provider and peer edges and $u_p$ and $u_r$ refer to the number of the $p$ and $r$ edges of $u$ respectively. We note that the cost function in Eq. 6.1 is intentionally made as simple as possible for two reasons. First, we want to concentrate purely on the consequences of our premises; thus, we avoid incorporating cost elements that can mask them. The second reason is simply analytical tractability. So basically, the first sum in Eq. 6.1 represents the most simple way of capturing our premises, and $\varphi_p$ and $\varphi_r$ are introduced for setting up a meaningful game (e.g., without attributing costs to the edges the game would end up in producing full graphs) but can be easily justified as inter-AS links clearly have maintenance costs. Also, note that we regard provider-customer edges to be financed unilaterally by the customer.

The Nash equilibrium of the Hierarchical Network Game (HNG) is a state such that no node can further reduce her costs by altering her strategy unilaterally. Since we have a network game, we will use the following more natural and slightly tailored equilibrium definition for our case:

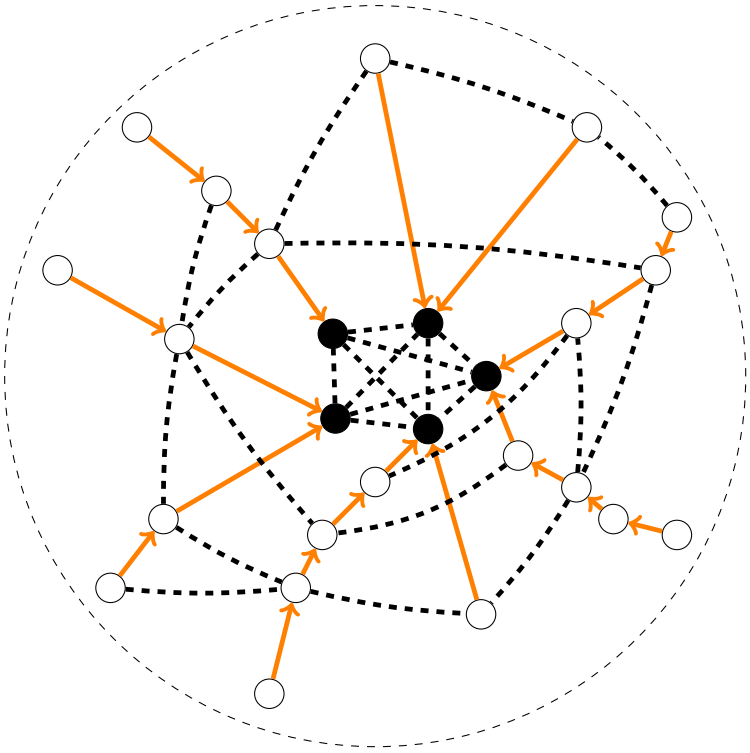


Figure 6.5: An example of the spider graph. The dashed and directed edges are the peer and customer-provider edges respectively and the black nodes are the ASes of the clique $\mathcal{K}$ i.e. the tier-1 ASes.

**Definition 2** (Pairwise Stable Nash Equilibrium (PNE)). *We say $G(s)$ constitutes a pairwise stable Nash equilibrium [88] if (a) Nash equilibrium, (b) $\forall uv \in E(G(s)) : C_u(s) \leq C_u(s') \wedge C_v(s) \leq C_v(s')$, where $s'$ differs from $s$ only in deleting $uv$ edge from $G(s)$, (c) $\forall uv \notin E(G(s)) : C_u(s) \leq C_u(s') \vee C_v(s) \leq C_v(s')$, where $s'$ differs from $s$ only in adding $uv$ edge to $G(s)$ and (d) contains no provider loops. Note that the latter requirement is fully in line with the Gao-Rexford conditions [68] ensuring BGP stability.*

### 6.1.3 Omnipresent subgraphs

Now we are interested in the equilibrium topologies of the HNG game as these structures will reflect the consequences of the function of the network, which is the provisioning of valley-free and local preference paths. For stating the claims, we need two more definitions.

**Definition 3** (Spider graph (Fig. 6.5)). *A graph is a Spider graph if it consists of:*

1. *a clique $K_r$ (representing the tier-1 ASes) comprising peer edges only*
2. *trees rooted at some subset of $V(K_r)$ having customer-provider edges, such that the provider in the relationship is always closer to the root than the customer*
3. *additional peer edges, such that $\forall \overline{uv}, \overline{uw} \in G(s) : t(v) \cap t(w) = \emptyset$, where $t(x)$ is the set of nodes in the subtree (i.e. the customer cone) of node $x$, including*

*x itself.*

**Definition 4** (Clear-cut Peer Edge (CPE))**.** *An $u\overline{v} \in G(s)$ edge is a clear-cut peer edge if:*

- $\varphi_r < \min\{\frac{|t(u)|}{N}, \frac{|t(v)|}{N}\}$
- $\nexists w \in \mathcal{P} : v \in t(w) \wedge u\overline{w} \in G(s)$.

Our first claim characterizes all meaningful states (i.e., where all the ASes can communicate with each other) of the HNG (and thus the AS topology) by identifying an omnipresent subgraph.

**Theorem 1.** *Every meaningful ($\sum C_u \neq \infty$) outcome of the HNG contains the Spider graph as a spanning subgraph.*

*Proof.* The subgraph of the customer-provider edges is a spanning DAG, as provider loops are not allowed. For having $\sum C_u \neq \infty$ the sinks of this DAG has to be connected by peer edges in pairs. Hence the set of the sinks correspond to the $K_r$ clique of the Spider graph.

Obviously each AS has a directed customer-provider path to some ASes of $K_r$. So one spanning forest of the DAG and the $K_r$ clique is a proper spanning Spider graph in the original graph. □

Using Theorem 1, we can characterize the pairwise stable equilibria of the HNG.

**Theorem 2.** *Every pairwise stable equilibrium of the HNG is the Spider graph.*

*Proof.* According to Theorem 1, any pairwise stable equilibrium contains the Spider graph as a spanning subgraph. Easily it contains a $K_r$ clique in which ASes do not have customer-provider edges. If there are any extra customer-provider edges, then there must be an AS which has at least two customer-provider links. Since the additional customer-provider edge does not reduce the communication cost but enlarges the maintenance cost, such an outcome cannot be a Nash equilibrium.

If the subgraph of the customer-provider edges is a forest, then in it if exists two nodes $v$ and $w$ such that $t(v) \cap t(w) \neq \emptyset$, then $v \in t(w)$ or $w \in t(v)$. Hence, if there is a peer edge $\overline{uv}$ such that there exists a node $w$: $\overline{uw} \in E(G)$ and $t(w) \cap t(v) \neq \emptyset$, then $v \in t(w)$ or $w \in t(v)$. Let $w \in t(v)$, so $u$ can reduce its cost removing $\overline{uw}$, which contradicts the definition of the Nash equilibrium. □

### 6.1.4 Placement of peer links

The following theorem gives a high-level insight into the placement of the peer edges.

**Theorem 3.** *If $G(s)$ constitutes a PNE then each peer edge is a CPE or part of $K_r$.*

*Proof.* We prove this indirectly. If there exists a peer edge out of $K_r$ which is not CPE then either $(i)$ $\varphi_r \not< \min\{\frac{|t(u)|}{N}, \frac{|t(v)|}{N}\}$ or $(ii)$ $\exists w \in V(G(s)) : v \in t(w) \wedge u\overline{w} \in G(s)$. For $(i)$ it is easy to see that at least for one AS it is worth to delete the edge. For $(ii)$ it's trivial that for $w$ is worth to delete $u\overline{w}$. In both cases we appear to a contradiction.

□

Finally our theorems lead to the following three corollaries.

**Corollary 1.** *In a PNE a peer edge appears only if it is in $K_r$ or its both endpoint ASes has sizable customer cones.*

**Corollary 2.** *For PNEs there exists an upper bound for the size of the customer cones of the ASes in $K_r$, or more formally PNE $\implies$ $\max_{u\in V(K_r)} t(u) \leq N(\varphi_p - \varphi_r(|V(K_r)|-1)+1)$.*

*Proof.* The cost of a node $u \in V(K_r)$ is $\varphi_r(|V(K_r)|-1)$. However, if $u$ leaves $K_r$ and creates only one customer-provider edge to another node in $K_r$, its cost would change to $\frac{N-t(u)}{N} + \varphi_p$. Hence in PNE

$$\varphi_r(|V(K_r)|-1) \leq \frac{N-t(u)}{N} + \varphi_p, \forall u \in V(K_r), \tag{6.3}$$

and thus

$$\max_{u\in V(K_r)} t(u) \leq N(\varphi_p - \varphi_r(|V(K_r)|-1)+1) \tag{6.4}$$

□

**Corollary 3.** *In the case of PNE there exists an upper bound for the size of $K_r$ independent from $N$, i.e. PNE $\implies$ $|V(K_r)| \leq \frac{\varphi_p+\varphi_r+1+\sqrt{(\varphi_p+\varphi_r+1)^2-4\varphi_r}}{2\varphi_r}$*

*Proof.* According to Corollary 2

$$\max_{u\in V(K_r)} t(u) \leq N(\varphi_p - \varphi_r(|V(K_r)|-1)+1), \tag{6.5}$$

and obviously

$$\frac{N}{|V(K_r)|} = \text{avg}_{u\in V(K_r)} t(u) \leq \max_{u\in V(K_r)} t(u), \tag{6.6}$$

hence

$$\frac{N}{|V(K_r)|} \leq N(\varphi_p - \varphi_r(|V(K_r)|-1)+1). \tag{6.7}$$

Dividing by $N$ and rearranging the inequality we get:

$$0 \leq -\varphi_r|V(K_r)|^2 + (\varphi_p+\varphi_r+1)|V(K_r)| - 1, \tag{6.8}$$

implying

$$|V(K_r)| \leq \frac{\varphi_p+\varphi_r+1+\sqrt{(\varphi_p+\varphi_r+1)^2-4\varphi_r}}{2\varphi_r} \tag{6.9}$$

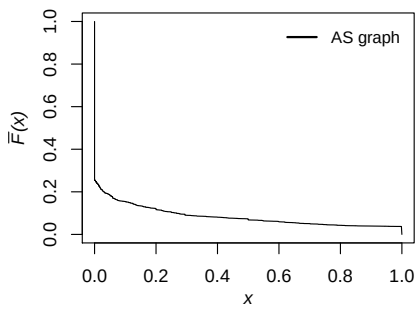
□

The above theorems deliver the following high-level sketch of the AS topology as a main intuitive message: (*i*) it is a Spider-like graph with a clique (of tier-1 ASes) in the center and trees routed in the nodes of the clique, (*ii*) the peer edges appear more likely between ASes having sizable customer cones, (*iii*) the size of the clique is constrained by the maintenance cost of peer and customer-provider relationships and (*iv*) the largest customer cone size in the nodes of the clique is also driven by these maintenance costs.

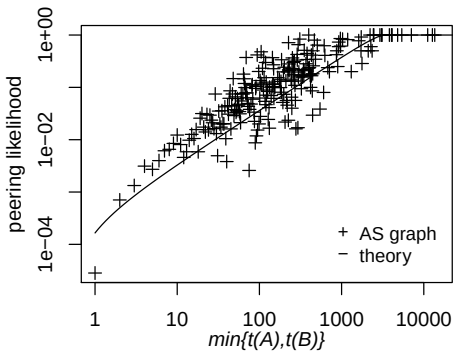
### 6.1.5 Discussion and double-checking against measurement data

For validating our analytical results, we used the AS Relationships dataset of May 2012, provided by CAIDA [33]. Although this dataset received some criticism over the last years, at this moment, no other sources of data are available containing more accurate tracing of the peer and customer-provider edges at the AS level.

This dataset contains AS-AS relationships for 41203 ASes with 57158 peer and 83374 customer-provider edges, thus let us build a labeled AS graph. Regarding Theorem 1 and 2 we investigated the existence of the Spider graph in two steps. First, we followed the customer-provider relationships in a top-down manner proceeding from the top tier-1 clique and kept all the nodes we could reach, this way we get a 92.5% node coverage which properly validates that the AS graph meets the first two properties (clique inside and trees rooted on the nodes of the clique) of Spider graphs. Secondly, we examined how typical for an AS $C$ with peering neighbors $A$ and $B$ that $t(A) \cap t(B) = \emptyset$. In other words, we calculated how typical is that the customer cones of the peers of an AS are overlapping (this is the direct checking of the third property of Spider graphs see Definition 3). For this, we randomly sampled the measured AS graph by choosing 500000 ($A$, $B$) node pairs for which $C\bar{A}, C\bar{B}$ exists. In each sample, we drew AS $C$ according to a degree-weighted probability function, and then we picked the peering neighbors with a uniform normal distribution. Our results confirmed that more than 75% of the pairs (Fig. 6.6a) have zero overlappings, and in other cases, the ratio of overlapping vanishes very quickly. These results readily support our claim that the AS level Internet topology is a Spider-like graph.



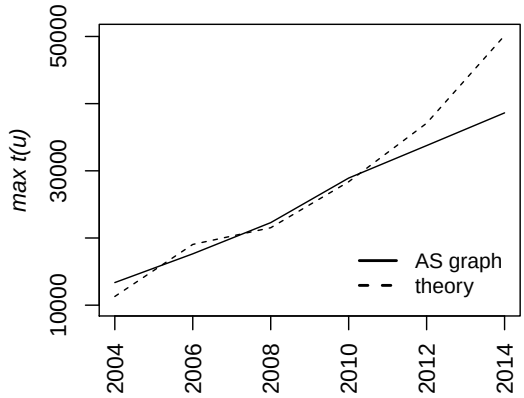

(a) CCDF for coverage overlapping of peer edges of an AS defined as $x = \frac{t(A) \cap t(B)}{min\{t(A),t(B)\}}$



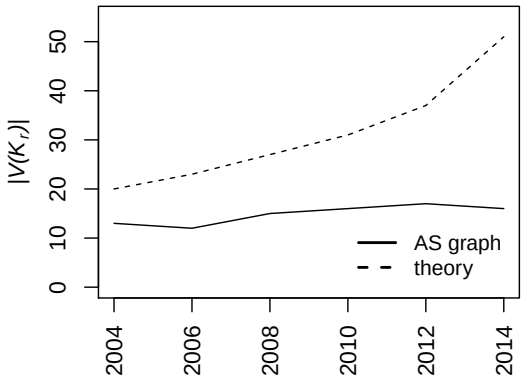

(b) Peering likelihood between ASes as the function of their customer cone size.

After that, as a next step, we measured the peering likelihood between two ASes as a function of the minimum of their customer cone sizes. The AS graph dataset of Fig. 6.6b shows the empirical probability that two ASes with a given minimum customer cone size ($\min(t(A), t(B))$) are in a peering relationship. The dataset supports that the peering likelihood is in a high correlation with the customer cone sizes of the ASes in the peering relationship.

Finally, we present a short argument illustrating our deductive predictions on the maximum customer cone size and the max size of the tier-1 clique. For doing




(a) Comparing our upper bound for $max\ t(u)$ based on Corollary 2 with the AS graph over time.




(b) Comparing our upper bound for $|V(K_r)|$ based on Corollary 3 with the AS graph over time.

this, we used historical AS datasets from CAIDA. Based on the number of customer-provider and peering relationships we have estimated $\varphi_p = \frac{Nc_1}{\#\text{of c-p edges}}$ and $\varphi_r = \frac{Nc_2}{\#\text{of peer edges}}$ with $c_1 = 1.1$ and $c_2 = 0.05$. Using these values, we have computed the results of our corresponding theorems and measured the maximal cone size and tier-1 clique size as a function of time in the CAIDA datasets. Fig. 6.7a shows that our rough estimation about the maximal customer cone size in the AS level Internet approximates the measured one based on CAIDA snapshots at a reasonable extent. Fig. 6.7b shows the prediction of our model regarding the size of the tier-1 clique. Although our simple formulae forecast a more increasing trend, the order of magnitudes is quite the same in both cases.

As a discussion, we first call the reader to notice the complementary nature of the deductive findings as opposed to the existing inductive models. While the existing inductive models concentrate on degree distribution, clustering, diameter, etc. the deductive reasoning give hints about spanning subgraphs, peering likelihood and constraints on the size of different parts of the network. We also recall that our deductive model is extremely simple and squeezes all maintenance cost related quantities into two constants $(\varphi_r, \varphi_p)$. In the light of this simplicity it is remarkable that the model gives practically usable predictions regarding the size of the tier-1 clique and the maximal customer cone of an AS.

One may argue that the results coming out of our deductive analysis are somewhat weak and don't say too much about the AS network. Such criticism may seem to be all right at first, but we find to be important and interesting in itself that the found topological peculiarities (summarized in Theorem 1,2,3 and Corollary 1,2,3) are direct consequences of the used BGP policies and thus will be present on the AS topology as far as these policies are at use. We believe that showing this causality contributes to our very limited amount of information about the Internet AS level topology. Finally, we note that more powerful premises can lead to more precise deductive topology characterization in future works.

## 6.2 The nature of the hierarchy in word networks

Everyday life is full of complex networked systems that humans recurringly navigate on a daily basis (e.g., traveling between locations in a city using public transportation). The available navigational datasets [119, 54, 171, 87] and models [93, 169, 157, 2, 26, 49, 171, 87] considering networked systems mostly target uncovering the average properties of a group of subjects and capture collective human behavior. Moreover, in terms of human navigation, the existing experiments focus on the dynamic process of learning to navigate, i.e., how people incrementally learn an approximate map of the network. Thus, existing datasets do not have sufficient data or appropriate tracing methods, permitting the analysis of long-term individual patterns. Here, we analyze the results of an experiment [98] with human subjects solving navigational tasks in a complex word-morph network. The recorded average of 40.9 timely ordered paths from 259 subjects and more than 200 paths from 9 subjects make the analysis of individual human navigation patterns possible. In contrast to existing studies, this amount of data enables the inference of characteristics about the steady-state way in which people choose a path between endpoints of a network *after* they have learned how to navigate the network in their particular way. We argue that this routine navigation process is more valuable to investigate since people use this approach on a daily basis. In the remainder of this text, we refer to this steady-state navigation simply as *human navigation*.

The nodes of the word-morph network, from which we process the navigation paths, are English words that are connected if they differ in only a single letter. In this vast and complex network, human subjects are given navigational tasks, i.e., to reach a destination word from a starting word by changing only one letter at a time, while still having meaningful words in intermediate states. Figure 6.8a shows a sample fragment of the word-morph network and two solutions (a shortest path and a human path) of the task with the starting word "yob" and destination word "way".

Network theoreticians across many disciplines[168, 170, 12, 124, 31] argue that the shortest path, i.e., the path containing the minimal number of intermediate steps in a network, between a source and a destination is a useful approximation of the real paths between them. In contrast with these results, our study shows that human subjects frequently apply detours, even in the long run. Our main finding is that these detours are the consequences of how an individual interprets a complex networked system on its level. We show that people tend to build up a significantly simpler representation of the word-morph network in the form of a hierarchy in their minds. A hierarchy is a way of interpreting an interconnection network by defining a central node (or a set of nodes) and referring to all other nodes with positions relative to, i.e., "above" or "below", the central node. These hierarchies are then used as helper structures when forming the paths in the network. In this work, we will refer to these hierarchical helper structures simply as "scaffolds".

As a result, real paths will be somewhat longer than the shortest alternatives, but the detours will be characteristic to the individual taking them, as no two individuals may abstract the same hierarchy of the network. Although existing models are assuming latent hierarchical scaffolds aiding navigation [169, 55, 94, 49, 68, 81], this is the first study processing sufficient individual human navigation data

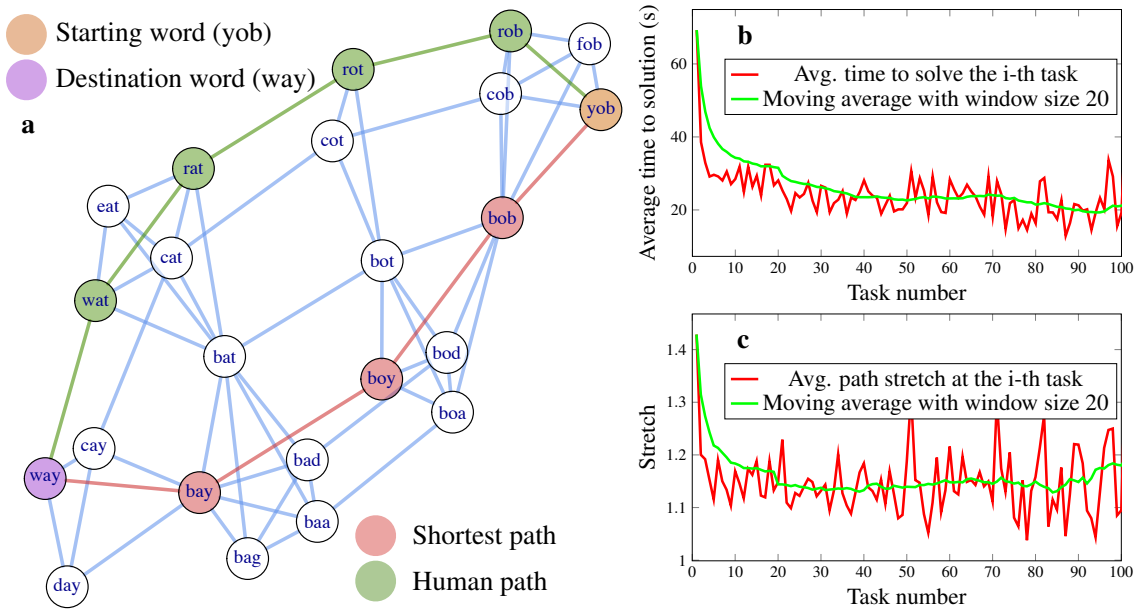



Figure 6.8: An example and high-level statistics of our navigation experiment. Panel (**a**) shows a sample section of the network of three-letter English words, in which two words are connected if they differ only in a single letter. When human subjects solve a navigation task, they come up with a path from a randomly given starting word to a destination word by changing only a single letter at each step such that they always obtain a valid intermediate English word. The red and green paths show a shortest and a slightly detoured human solution from "yob" to "way". Panel (**b**) presents the average time it takes for human subjects to solve the n-th task in a row, while panel (**c**) shows the stretch of the human paths, i.e., the ratio of the length of the paths found by human subjects to the length of the shortest possible path in the word-morph network. While the average time to solve a task clearly decreases with the number of tasks solved, the stretch of the solutions stabilizes between 1.2 and 1.1. This suggests that human subjects develop a specific strategy in the first few rounds, but after a few tens of solved tasks, their strategy is not improved any further in terms of length. Therefore, they have a simplified interpretation of the network, and they find their paths through this, only slightly faster as time elapses.

to visualize and analyze these individually created hierarchies.

We discuss that navigational scaffold hierarchies may boost the learning process to navigate the word-morph network and reduce the memory requirement of navigation by order of magnitude. Moreover, identifying the individual scaffold hierarchies as the enablers of memory-efficient navigation in the word-morph network is of particular importance since this may promote uncovering of navigational schemes in other complex networked systems considering not only humans. Similar detours have been identified in measurements capturing the collective behavior in networks from diverse areas of life. Gao et al. showed that the paths of packets going through the internet are also detoured to a non-negligible extent [69], and they showed that the hierarchical policies of internet packet routing may be responsible for a major proportion of inflation. Detours have been identified in road networks by Zhu et al. [177] and in cattle pen systems by Grandin [73], while similar phenomena were also

reported in airports [49, 150] and brain networks [8, 49].

### 6.2.1 Results

For our research, we use data from an experiment with a word-morph game application for smartphones [65] (see Methods for details). The application collected 19828 paths from 259 human subjects navigating the word-morph network, and the corresponding dataset was published in Scientific Data [98]. After cleaning the data from paths not referring to steady-state navigation, by removing tasks that were either unfinished, contained loops or took an extraordinarily long time ($> 300$ seconds) to complete, our working dataset of paths was reduced to 10857 paths (for more details about data filtering, see Methods). The word-morph network is a complex network that is impossible for a human subject to keep fully in mind with its 1008 nodes and 8320 edges. The values of the average degree (i.e., the average number of edges emanating from the nodes), the diameter (the longest shortest path in the network) and the clustering coefficient [170] of the network are 16.39, 9 and 0.44 respectively. To attain a high-level impression about the performance of human navigation, we have plotted the average time needed to solve the n-th task in a row in Figure 6.8-b. We can see that after a few initial rounds, human subjects find a solution in approximately 30 seconds on average, and from there on, they slowly improve to approximately 20 seconds after solving 100 tasks. Notably, it is an intrinsically astonishing finding that after a few rounds, people can find paths in this complex maze very efficiently. Strikingly, the improvement in time does not imply that the paths found are also shorter. In Figure 6.8-c, the stretch of human solutions is shown compared to the shortest paths. The stretch of a path $P$ is computed as the ratio of the length of $P$ to the length of the shortest path between identical starting and destination words. In the example of Figure 6.8-a, the stretch of the human path (green) is $\frac{5}{4} = 1.25$ compared to the shortest possible path (red). Figure 6.8-c shows that although human subjects improve in terms of the time needed to solve a task, the stretch of the paths they find stabilizes slightly below 1.2. Thus, the length of the human paths seems not to converge to the length of the shortest path (i.e., to stretch 1), and they always include some detours. A plausible explanation for this is that human subjects develop some sub-optimal strategy through the course of the game and use this strategy to solve upcoming tasks. The improvement in time only means that the application of the same strategy becomes increasingly more effective. Nevertheless, how can we characterize the strategy in use?

Panels **a** and **b** in Figure 6.9 illustrate how differently an algorithm implementing shortest paths and a single human subject use the word-morph network to solve the navigational tasks. The plots show only edges traversed more than two times in the course of solving 1000 tasks. In the case of the shortest path algorithm, the usage of edges is homogeneous. The algorithm has no clear concept or in-depth interpretation of the word-morph network. It thus picks the paths mechanically without any sign of favoring specific regions of the network. The selected human subject behaves quite differently. The subject seems to have a clear concept of the network. The subject structures the network in a subjective manner by identifying various regions and places a larger emphasis on nodes and edges connecting these regions. A clear sign of this structuring is that from the human solution, a hierarchical scaffold structure

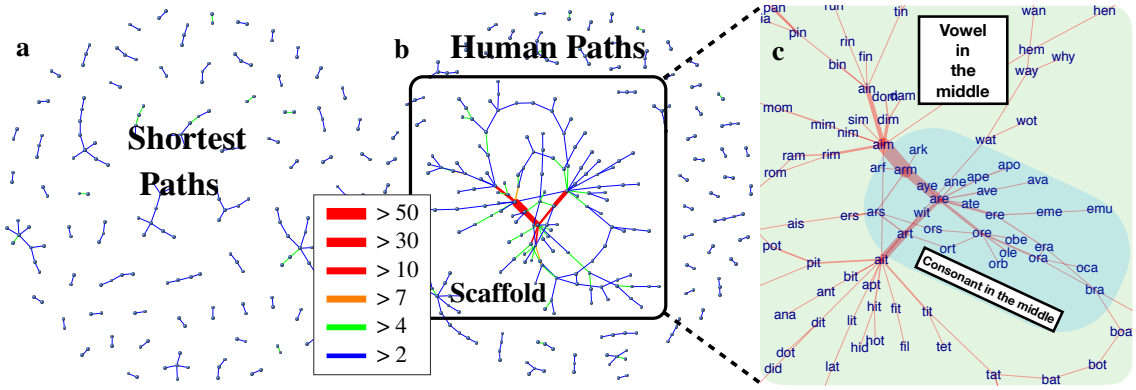



Figure 6.9: Structures behind human paths and shortest paths. Panel (**a**) shows how many times an edge is crossed after solving 1000 random tasks by using the shortest path between the source and target word. The almost homogeneous distribution of edge crossings suggests that the entity using these paths does not have any form of understanding or interpretation of the word-morph network; conversely, it mechanically picks paths. Human paths are quite the contrary. Panel (**b**) shows the edge crossings of a single human subject when solving the same 1000 random tasks. The human solution appears to be highly structured, suggesting that humans possess a characteristic concept of the word-morph network. The structure is very close to a pure hierarchy. There is a clear scaffold that guides navigation, consisting of red, orange, and green edges with a high number of crossings. This scaffold shows that the human subject tends to simplify the problem and form a simpler and systematic, although not necessarily optimal, strategy. From the sides of the network, where a navigation task starts, the human subject tends towards the scaffold where a switch is performed to other sides of the network. How this particular scaffold is built up is quite specific. Panel (**c**) shows the words in the middle of the scaffold. "Aim", "art", "arm" and "are" depict words where consonants and vowels can be changed very effectively. In this case, the scaffold is used to switch between regimes of the network based on the location of vowels and consonants.

is formed (see Figure 6.9-b for an example). To capture this behavior, we focused on subjects highly engaged with the game, thus producing enough data to examine the navigation strategy they use deeply. We investigated subjects having more than 200 completed navigation tasks (9 subjects qualified for this). For these subjects, we processed all the solutions of the navigational tasks and assigned weights to the edges of the word-morph network, reflecting how many times they were used in the solutions. We dropped the rarely used edges, for which the usage could be the result of randomly choosing the source and destination words. From the remaining graph, we took the largest component as the scaffold. In 90% of the cases, the scaffolds of the human subjects were at least two times larger in size compared to the random case, but in the majority of the cases, the human scaffolds were found to be an order of magnitude larger (see Panel **a** in Figure 6.10).

Panel **b** of Figure 6.10 shows that the average degree of the scaffolds is approximately 2 in the case of all subjects. This means that the scaffolds are tree-like connected sub-networks of the original word-morph network. This result is fully in line with the assumptions of existing hierarchical human navigational models[169,

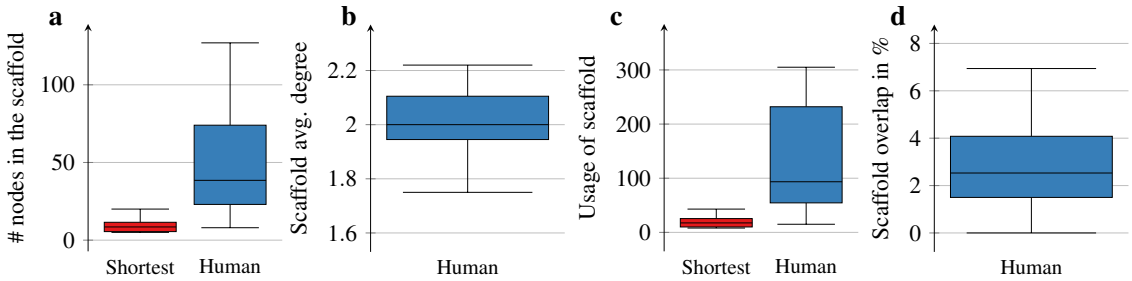



Figure 6.10: Properties of individual human scaffolds. Panel (**a**) shows the size of the human scaffolds compared to the shortest path case. The human subjects' behavior clearly deviates from the shortest path algorithm, as they form sizeable navigational scaffolds compared to shortest paths. The average degree of the scaffolds is close to approximately 2, as shown in panel (**b**); thus, the structure is very close to trees. Panel (**c**) confirms that the scaffold is heavily used by human subjects when completing the navigation tasks. We define usage simply as the sum of intersections between the subject's paths and the scaffold. If we denote the solutions of the subject as $P_1, P_2 \dots P_K$, where $K$ is the number of puzzles solved by the subject, then the usage of the scaffold $S$ is computed as $\sum_{i=1}^{K} E(P_i) \bigcap E(S)$, where $E(P_i)$ denotes the set of edges contained in $P_i$, while $E(S)$ is the set of edges in the scaffold. Panel (**d**) shows that the individual human scaffolds are indeed "individual," as the observed overlaps between the subjects' scaffolds are only 2.6% on average.

55, 94, 81]. Compared to the shortest paths, the edges of the scaffolds are heavily used by the subjects (see Figure 6.10-c) with a very specific usage pattern. The scaffold has a definite core of a few nodes, between which the usage of the edges can exceed 50 in the particular example of Figure 6.9-b. This core behaves as a switching device among different parts of the network and abstracts the individual's concept of the structure of the whole network. The scaffold is built up in a hierarchical, tree-like fashion, as edge utilization drops when receding from the core. In the course of navigating between words, subjects use the scaffold as a guiding framework. Figure 6.9-c shows the words residing in the scaffold. In this example, the network is clearly divided into regions based on the position of consonants and vowels in words, and the core words are picked by the human subject in order to switch effectively among these regions. Our results show that although these individual scaffolds may have some similarities, every subject used a fairly unique set of nodes and edges forming their hierarchical scaffolds (see supplementary Figure 1 for additional examples of personal scaffolds). This finding is readily supported by Figure 6.10-d, which shows the percentage of overlap between all possible pairs of scaffolds. The overlap for scaffolds $i$ and $j$ is computed according to the Jaccard index over the sets of edges: $\frac{E(S_i) \bigcap E(S_j)}{E(S_i) \bigcup E(S_j)}$, i.e., the ratio of edges present in both scaffolds ($E(S_i)$ denotes the set of edges contained in scaffold $i$) to the edges in the union of the scaffolds. Thus a network's overlap with itself is practically 100%. One can see that in the case of the scaffolds of the subjects, the average of the overlap is minimal, approximately 2.6%, and the maximum overlap is only 7%.

To quantify the statistical significance of the results regarding the scaffolds, we tested the null hypothesis that human paths can be explained by the shortest path algorithm. To test this hypothesis, we generated 500 solutions with the random

| Parameters of fitting and p-values for scaffold sizes | | | |
|---|---|---|---|
| # | Wei. shape | Wei. scale | p-value |
| 1 | 3.06 | 25.22 | 4.57E-62 |
| 2 | 3.09 | 12.83 | 0.00E+00 |
| 3 | 3.83 | 18.97 | 3.10E-03 |
| 4 | 4.14 | 16.38 | 3.56E-04 |
| 5 | 3.81 | 14.50 | 6.98E-149 |
| 6 | 4.07 | 5.16 | 2.61E-03 |
| 7 | 3.96 | 9.37 | 1.70E-304 |
| 8 | 3.26 | 9.99 | 3.05E-04 |
| 9 | 4.25 | 7.36 | 0.00E+00 |
| Parameters of fitting and p-values for scaffold usages | | | |
| # | Wei. shape | Wei. scale | p-value |
| 1 | 2.92 | 87.47 | 9.58E-202 |
| 2 | 2.86 | 25.26 | 0.00E+00 |
| 3 | 3.72 | 40.97 | 1.99E-03 |
| 4 | 4.19 | 38.74 | 4.78E-05 |
| 5 | 3.51 | 28.97 | 0.00E+00 |
| 6 | 3.18 | 8.60 | 2.84E-03 |
| 7 | 3.50 | 18.60 | 4.31E-298 |
| 8 | 2.89 | 19.50 | 1.05E-04 |
| 9 | 3.93 | 14.91 | 0.00E+00 |

Table 6.2: Statistical analysis of scaffold size and usage. The null hypothesis is that the solutions of human subjects are random shortest paths. To test this hypothesis, we generated 500 solutions with the random shortest path algorithm over the same set of puzzles that the subjects solved. Parameters of the Weibull distributions fitted to the scaffold sizes (left panel) and usages (right panel) and the p-value referring to the null hypothesis are given for all the subjects.

shortest path algorithm over the same set of puzzles that the subjects solved. We found that the distribution of scaffold sizes and usage can be nicely estimated with a Weibull distribution (see Methods) in the case of all subjects. Table 6.2 shows the parameters of the Weibull distributions fitted to the scaffold sizes and usages plus the p-value indicating the tail probability that a scaffold of similar size and usage to the human solution could be derived from randomly chosen shortest paths. The p-values never exceed the alpha level of 0.05 and are extremely small in most of the cases, meaning that we have to reject the null hypothesis with high statistical significance. This substantiates the conclusion that the behavior of the human subjects cannot be explained based on the shortest path algorithm.

The identification of the individual scaffold hierarchies as core switching devices in the human interpretation of the word-morph network poses an intriguing question: Why do we use them even after mastering our ability in the navigation task? Why do we tolerate sub-optimal paths through these scaffold hierarchies and not strive for shorter paths? Recall that detours in the subjects' paths persisted even after completing 100 navigation tasks. We argue that the reason behind this is related to our information encoding and processing capabilities. In short, we build scaffold

hierarchies while being satisfied with sub-optimal paths because this way, we do not have to process every bit of information about a large and complex system, and we can get away with an interpretation that is an order of magnitude simpler. To show this, we use the following minimalist information-theoretic model inspired by our results above. The word-morph network is represented by a graph $G(N, E)$ defining its nodes $N$ and edges $E$.

For modeling human behavior, we use a simple tree hierarchy as a scaffold for navigation. The construction of the hierarchy proceeds by picking the node with the highest closeness centrality [18] and building the breadth-first search (BFS) tree emanating from it. This BFS tree will be used as the scaffold. Inspired by the information exchange algorithm well-fitted for hierarchically structured organizations [55], we define human navigation based on the scaffold hierarchy as follows: $(i)$ if the destination node is below the current node or its neighbors in the hierarchy, then we step to its closest superior or the destination itself provided that the destination and the current nodes are connected; $(ii)$ if the destination node is not below the current node in the hierarchy, then we step to the current node's direct superior in the hierarchy. As an analogy, this simple navigation mechanism captures that if somebody is my subordinate in the hierarchy or the subordinate of someone that I know, then I know who is the closest to them among my acquaintances. If I know nothing about the target, then I turn to my direct superior. Note that this straightforward process gives only a possible way of using a very artificial scaffold, the BFS tree. Our goal with analyzing this simplified navigation process is to enable the information-theoretic analysis of the paths formed by the usage of scaffolds. Paths emanating from this simple model will clearly not match the paths used by any of the subjects for multiple reasons. First, although scaffolds built by humans are very similar to trees, they are not trees in many of the cases (see Fig. 6.10b). Second, human scaffolds vary subject by subject and have only a minimal overlap across subjects (see. Fig. 6.10d) and with the BFS hierarchy.

To characterize the complexity of implementing the paths provided by the shortest path algorithm and human navigation, we approximate the required minimum information in every node to decide which next step to take towards all destinations in the word-morph network. Let us assign positive integers, i.e., 1,2,3 ..., as IDs to the nodes of the network. At each node $x$, we can represent the amount of information needed to make the right choice by a node table $T_x$. At node $x$, this node table has $|N|-1$ entries (where $|N|$ is the number of nodes in the word-morph network) belonging to all the nodes other than $x$, and each entry contains the ID of a neighbor to take next towards a given destination. For example, a node table $T_5 = (1, 2, 1, 2)$ tells us that at node 5, if we want to go toward node $1, 2, 3, 4$, we should take nodes $1, 2, 1, 2$ as next steps, respectively. This node table implicitly tells us that node 5 is connected to nodes 1 and 2 and that, in this example, the network has five nodes. Now, tables $T_x, \forall x \in N$ contain all the information required to implement the given paths between arbitrary pairs of nodes in the word-morph network. To approximate how many bits of information are needed to store these tables in memory, we compute their empirical Shannon entropy[154], defined as $H_0(T_x) = \sum_{c\in\Sigma} \frac{n_c}{n} \log \frac{n}{n_c}$, where $\Sigma$ denotes the set of different numbers in $T_x$, while $n$ and $n_c$ represent the length of $T_X$ and the number of occurrences of each number $c \in \Sigma$ in $T_x$, respectively. Then, $\frac{\sum_{x\in N} H_0(T_x)}{n}$ yields the global per node entropy to

implement the paths.

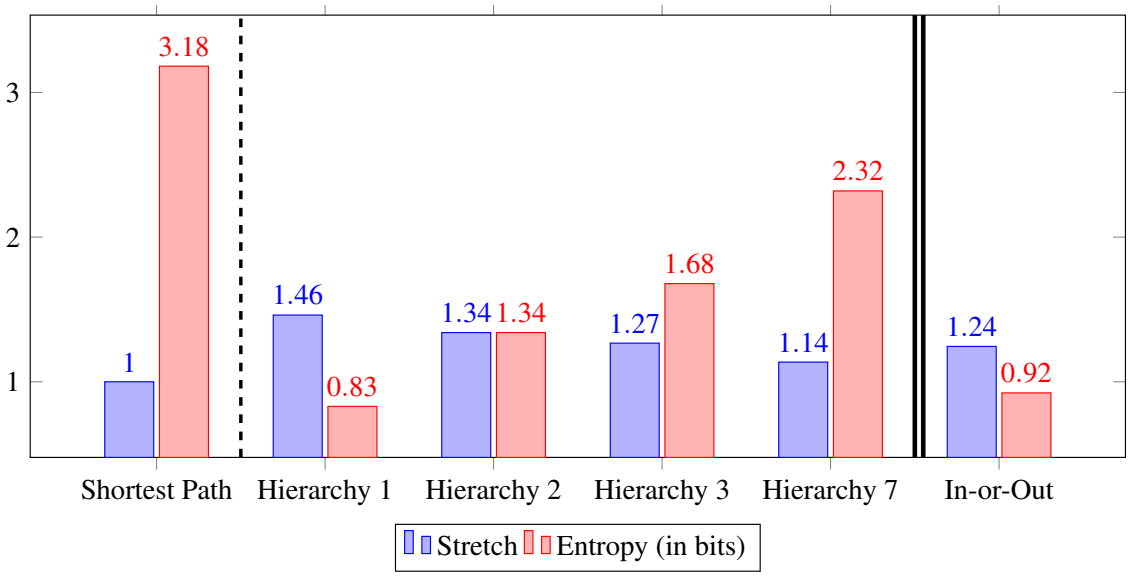



Figure 6.11: Comparison of stretch and entropy of various paths. Shortest paths clearly have a stretch of 1, but this optimality comes at a price of high entropy, i.e., a high memory requirement for storage. Hierarchies 1-7 show the very efficient stretch-entropy tradeoff if we memorize only 1-7 uplinks in the simplest BFS hierarchy. The decentralized In-and-Out hierarchy with one direct superior, based on the highest closeness centrality, is a sweet spot in this tradeoff space. This simulates the case when people know all subordinates in the network but remember only one superior closest to the center of the network. It provides a realistic stretch, but the required entropy is an order of magnitude lower than that in the shortest path case.

In Figure 6.11, the required information for implementing the shortest paths and hierarchical paths in the word-morph network are shown. Shortest paths have a stretch of one, but the price of this is high entropy, as approximately 3.18 bits per node are required to store the shortest paths in the node tables (see the Shortest Path column on the left of Figure 6.11). Navigation with the simple BFS scaffold has an order of magnitude less (approximately 0.83 bits per node) entropy (see the Hierarchy 1 column of Figure 6.11), but hierarchically guided paths are much longer; they have a stretch of 1.46. Recall that our results with human subjects indicate a stretch slightly below 1.2. Hierarchy 2, 3 and 7 columns in Figure 6.11 stand for a slightly modified version of the BFS hierarchy in which we do not have strictly one direct superior but can have links to at most 2, 3, and 7 superiors in the BFS tree, respectively. These hierarchies are no longer trees, but they are still as sparse as the human scaffolds. These modifications readily illustrate that there is a clear tradeoff between stretch and entropy. Having to remember more superiors reduces the stretch but inevitably increases the complexity. Nevertheless, with hierarchy 7, a stretch of 1.14 is achievable at the cost of only 2.32 bits of memory per node. These results readily illustrate that even the most rudimentary scaffold guiding navigation can achieve an effective stretch-entropy tradeoff. However, BFS scaffolds are constructed in a centralized fashion and rely on global information about the network, which is not realistic. A more realistic decentralized scaffold with only one

direct superior yields a sweet spot in this tradeoff space while being computable with local algorithms [175]. In this hierarchy, called In-or-Out, every node's superior is the neighbor lying in the most central location in the network in terms of closeness centrality. This simple, local strategy can provide a very low stretch for an order of magnitude less entropy compared to the shortest paths. This is because the In-or-Out hierarchy is aware of the neighbors' centrality; thus, every node's direct superior is a neighbor that is closest on average to any other node in the network. Interestingly, the In-or-Out hierarchy stretch is close to what we have observed with human subjects.

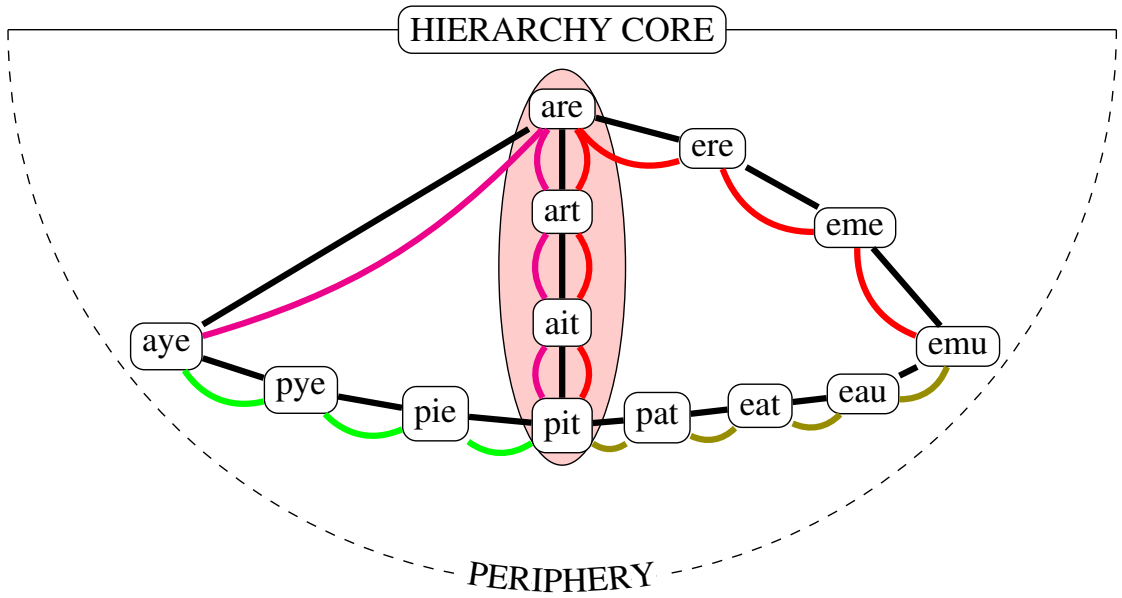



Figure 6.12: Shortest and hierarchically guided paths in the word network. Learning only the shortest paths between the words *pit* and *aye* and between *pit* and *emu* makes us conclude that the word *aye* is 7 nodes away from *emu*. However, with a hierarchical scaffold, a four-node path between *aye* and *emu* can be found even though both of the paths between *pit* and *aye* and between *pit* and *emu* are longer through the scaffold than through the shortest possible path.

In addition to simplifying the process of navigation, scaffold hierarchies can boost learning the structure of an unknown network by observing its paths. To show this, we use a straightforward incremental model where, in every step, we show a single path connecting randomly chosen nodes and compare the reconstructed network structure and the efficiency of navigation based solely on the given paths to the original network. Fig. 6.12 illustrates the steps of this learning process for the cases in which we show paths according to shortest or hierarchical scaffolds from the word-morph network. In the first case, we show the shortest paths between the words "aye" and "pit" (green) and between "pit" and "emu" (olive), and based solely on this knowledge, one may implicitly deduce a path from "aye" to "emu" traversing six nodes. Alternatively, showing paths using hierarchical scaffold yields somewhat longer paths (red). However, one can see that the newly gained path between "aye" and "emu" leads to a substantially shorter path requiring only three intermediate nodes. In Fig. 6.13, the integrity and the stretch and entropy footprint

of the various learning scaffolds are shown when we continue simulating the learning process up to 2000 paths with a computer program (see Methods for details). In panel (**a**), the size of the giant component in the network reconstructed from the paths is shown as a function of learned paths. The shortest path scaffold provides only very sporadic knowledge about the network in the initial $(0 - 120)$ learning steps, as the size of the giant component hardly grows with the number of learned paths. The most integrated knowledge is provided by the most simple scaffold of Hierarchy 1. In panel (**b**), we can clearly distinguish between two phases of the learning process. Until approximately 700 paths, rough exploration of the nodes and possible connections in the network occurs. According to the inset of the panel, by the end of this exploration phase, one can connect more than 90% of all possible node pairs in the case of all scaffolds. Using the shortest paths as learning scaffolds, we can find only very long paths in the exploration phase, as the average stretch can exceed even 3. Interestingly, if paths are picked according to a hierarchical scaffold, we can obtain paths with a lower stretch as the scaffold becomes increasingly simpler, i.e., the number of direct superiors decreases. In the case of the simplest one-superior case, the stretch is very stable at approximately 1.5. Therefore, in the exploration phase, one can learn reasonable paths much faster if paths are given according to a hierarchical scaffold. After the exploration phase, we do not explore new territories of the word-morph network; what we do is only improve our knowledge. In this improvement phase, the shortest path scaffold takes the lead over the hierarchical scaffolds, yielding the best stretch values. The price of being better in stretch is a higher entropy, as can be seen in panel (**c**): The entropy of the scaffolds is similar in the exploration phase; however, as the number of paths learned increases, the entropy of the simplest Hierarchy 1 scaffold starts to decrease substantially, while the shortest path one continues to increase almost linearly.

### 6.2.2 Discussion

Although this study concentrates on a networked system, the underlying problem of human navigation in the word-morph network seems even more interesting in light of the fact that the current explanations of physical navigation tend to apply models considering the graph-like abstraction of the surrounding physical environment. In fact, there is an ongoing debate about whether we build a detailed cognitive map or a much simpler cognitive graph of the possible physical choice points [129, 42] inside our head. Furthermore, recent studies reported major correlations between the navigation and learning skills of humans [128, 122], while others went even further and investigated the possibility that navigation in cognitive spaces may lie at the core of any form of organized knowledge and thinking [20, 58, 19]. The word-morph network is a special mixed system over which navigation relies strongly on domain-general mechanisms since both spatial, manifested in the Hamming distance between words, and cognitive, i.e., the function and meaning of the words, dimensions contribute to the formation paths. Thus promising speculation is that the identification of individual scaffolds guiding human navigation in the word-morph network may contribute to a better understanding of how humans structure, encode and navigate through cognitive spaces.

The empirical confirmation of individual scaffold hierarchies may also, help re-

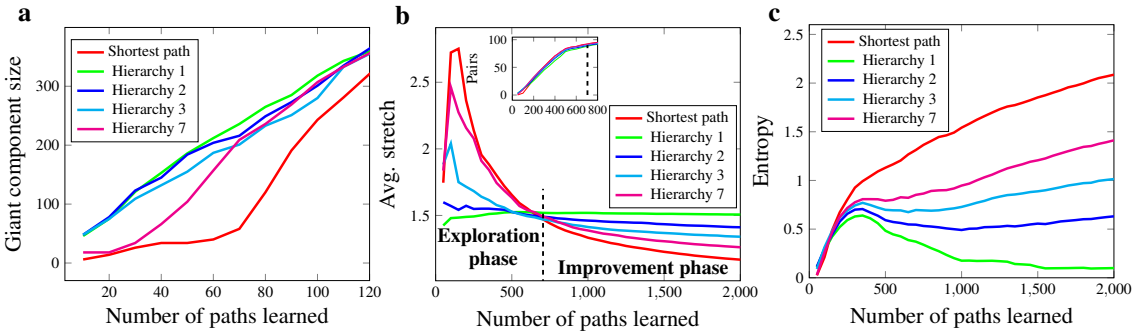



Figure 6.13: Learning curves in the word-morph network. Panel (**a**) shows the size of the giant component vs. the number of paths learned according to various learning scaffolds. Using the shortest paths as the scaffold yields sporadic knowledge about the network, especially in the initial steps of learning, since the size of the giant component is very low compared to the other scaffolds. The most integrated knowledge about the network is given by the simplest Hierarchy 1 in the initial steps of learning. The inset of panel (**b**) shows that after learning only approximately 700 paths, one can infer valid paths between 90% of all possible node pairs using either the shortest path or hierarchical scaffolds. In this exploration phase, learning based on shortest paths seems to be quite inefficient, as the stretch can even reach 3. In this phase, the simplest hierarchical scaffold yields the shortest established path on average. Only in the improvement phase, in which no significant new parts of the word-morph network are explored, is the relation is reversed. The entropy of the paths is shown in panel (**c**). The exploration phase shows no difference among the scaffolding schemes; however, in the improvement phase, the entropy of the hierarchical scaffolds is much lower compared to the shortest paths.

solve known anomalies in modeling human navigation behavior in networks. Human paths over networks are reported to exhibit non-negligible memory [150, 153, 158], which leads to problems when applying first-order Markov chains to approximate paths in spreading dynamics and community detection [150]. Individual scaffold hierarchies explain the source of these anomalies, as the next step of hierarchically guided paths depends on nodes visited previously by the given individual. Building on the assumption of hierarchical scaffolds behind network paths, we may be able to refine higher-order Markov models, which may bring us closer to a better understanding of how real systems are organized and function.

### 6.2.3 Methods

**Dataset** – For our study, we have used the dataset collected by a smartphone application called "fit-fat-cat" running on the Android platform. The dataset [98] is published in Scientific Data, with the appropriate ethical consent. Here, we summarize the data collection process; for a detailed description of the experiment, consult [98]. The application is available from the Google Play store [65]. When a subject starts a navigational task, the source and destination words are generated randomly from all possible three-letter English words. The source and destination words are displayed in a box (see Figure 6.14). Below this box, the list of words that the subject visited so far in that particular task is shown. When starting a new

task, the list contains only the source word. The subject can enter the consecutive words in a user-friendly manner by using a virtual keyboard of the phone. First, the subject selects the letter to change, then chooses the new letter with the keyboard. After changing a letter, the app automatically adds the new word to the list. In this way, the subjects can see which words they have already tackled when solving a particular navigation task. A task may end in three ways. If the subject reached the target word through such one-letter transformations, then the task is solved. In this case, the word becomes green-colored to show the end of the task. Second, the subject can give up the task by pressing the "new game" button. In this case, the subject acquires the next task automatically. Finally, the subject can press the "magic wand" button. In this case, a possible (shortest path) solution of the task is shown before starting a new task. No matter how the task is ended, the list of words is anonymously submitted to our database stored in the cloud. Due to the scale of the experiment, we couldn't control the external conditions under which the subjects carried our the solutions, apart from standard software checking of the validity of the subjects' inputs. For more details, see [98].

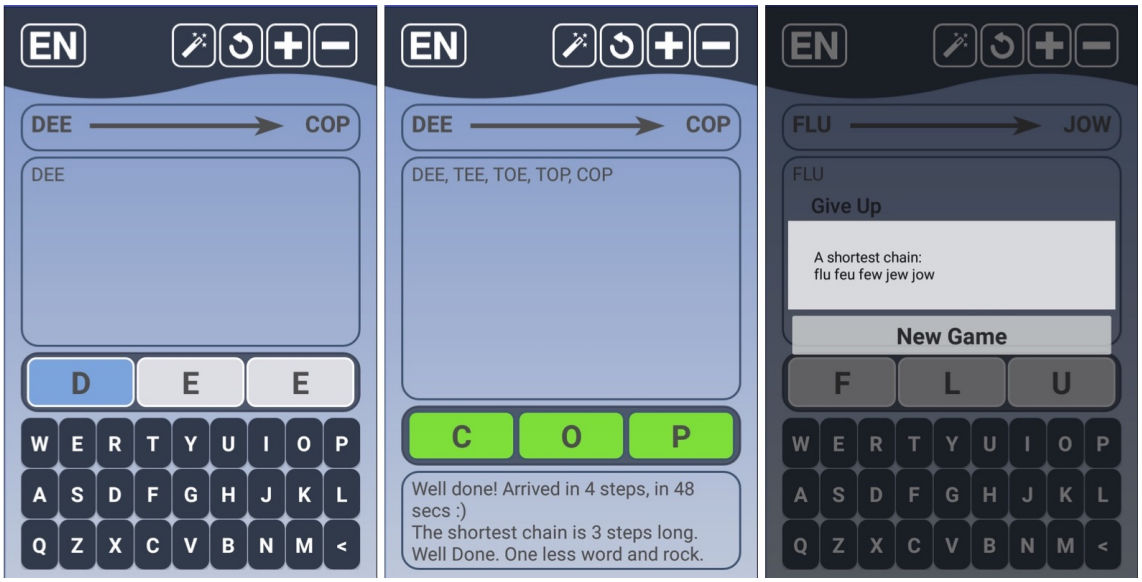



Figure 6.14: The main screen of the fit-fat-cat application.

Detecting an individual scaffold requires a relatively high number of completed navigation tasks. Completing many puzzles can be a very tedious and repetitive task. Doing this in a single row (e.g., in a paid, controlled experiment during which the subject can concentrate from the beginning to the end) is arguably unfeasible. Luckily, 9 of the subjects found the game interesting enough to solve more than 200 puzzles. Thus it is not the number of subjects that are uniquely large in the dataset, but the number of paths collected from a single subject.

**Path filtering** – Instead of focusing on the dynamic process of how we learn to navigate, i.e., how we learn an approximate picture of the network by exploration, we concentrate on the way people routinely choose paths in a network $after$ they have developed an individual path selection strategy. In this steady state, subjects do not explore the network or wander around; they solve the puzzle by routine. To analyze this steady-state behavior, we have to drop all unfinished paths, paths

taking too much time to complete and loops from the dataset. Of the recorded 19828 paths, we dropped 8177 because they did not reach the target for some reason, 712 paths because the time to solve the puzzle was unusually large ($>$ 300 seconds), which raises the question of if the subjects concentrated on the puzzle, and only 352 paths (1.7% of the total paths) because they contained loops.

**Weibull fitting to the random shortest path algorithm** – The scaffold sizes and usages of the random shortest path algorithm can be well-estimated with a two-parameter Weibull distribution. As an illustration, we verify the goodness of the fit for the puzzle set of subject 4 in Fig. 6.15. The results for the other subjects are highly similar.

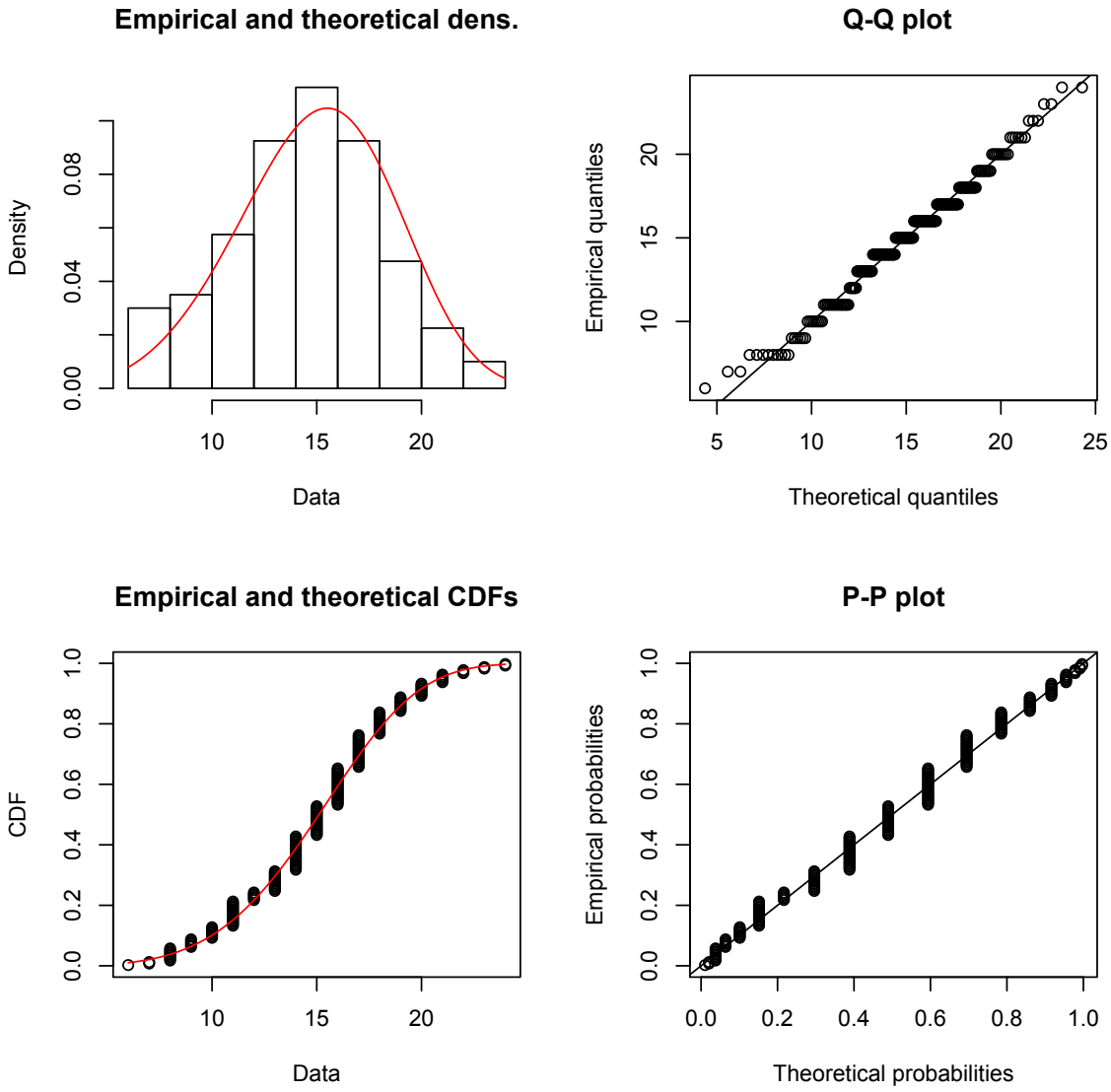



Figure 6.15: Goodness of fit of the Weibull distribution to the scaffold sizes given by the random shortest path algorithm over the puzzles of subject 4.

**Computer simulations** – For the investigation of the incremental learning of a network via its paths, we have written a simulator in the Python programming language. In the beginning, the simulator reads the network $N$. After that, it iteratively picks random pairs from the network and computes the shortest and

hierarchical paths between them according to the given BFS hierarchy. At each iterative step, the current knowledge about the network is the union of nodes and edges contained in the previous iterations. Therefore, at step $t$, the knowledge about the network is a graph $G_t(V, E)$; then, after adding a path $P_t$, it is extended to $G_{t+1} = G_t \bigcup P_t$. The simulator computes the required entropy and stretch of the paths in $G_t$ compared to the shortest paths in $N$ every 50 steps. We note that we have run the simulations beyond 2000 paths, but the relative positions of the stretch and entropy plots of the algorithms remain the same in that regime.

**Data Availability** – The data supporting the findings of this study are available from the "fit-fat-cat" public Open Science Framework data repository [99] and described in detail in [98].

# Chapter 7

# Conclusion

Which was first the chicken or the egg? In this dissertation, we have studied a question very similar in spirit. What is first, the network, or the paths? Our work points out that there is a compelling co-evolution between the network structure and the operational paths taken by entities using the network for transferring many kinds of information. We have seen that operational paths are not necessarily shortest paths. Paths are thus not mechanical results of an algorithm working over the network. Real paths are the results of a non-trivial path selection process serving as the main function of the network. We have studied the consequences of two classes of path selection hypotheses on the structure of the network by using a new approach called function→structure analysis.

The function→structure analysis of navigable networks, which are navigable by distributed greedy search algorithms guided by the metric space underlying the network, yielded a whole new way of generating realistic complex networks. Formalizing the function→structure analysis as a game-theoretical problem, we have been able to characterize the properties of networks, ensuring maximal navigability in hyperbolic and euclidean spaces. The hyperbolic navigation game results yielded very realistic, complex network structures with a scale-free degree distribution, small diameter, and high clustering coefficient. Interestingly, applying the model on the three-dimensional coordinates showing the location of various parts of the human brain obtained by MRI, yielded edges present in the real human brain with very high probability, in the form of dense neural connections. On top of the basic structural features, our analysis located the edges critical for navigability. The collection of critical edges constitute the greedy frame, which has to be included as a subgraph in any network ensuring maximal navigability over a given metric space. We have shown that adding a small subset of the greedy frame, missing from a real navigable network, can dramatically increase the navigability of the original network.

Numerous real networks exhibit hierarchical structure, and the connections between the nodes reflect relationships concerning the hierarchy. Social and organizational networks are clearly hierarchical, but the most pathological example of hierarchical networks is the Internet, which is the interconnection system of internet providers and customers. The path selection process on the Internet is also formulated in terms of the customer-provider hierarchy. Our function→structure study of the Internet revealed a complementary insight to hierarchical complex networks. We have shown that the path selection function used on the Internet requires a well-

defined subgraph to be present to provide policy-compliant paths between arbitrary Internet domains. Our analysis also explained some rules of thumbs. The peering likelihood and the possible customer cone sizes of peering ASs are well-known by Internet practitioners. Still, this work proves a theoretical connection between the applied path selection policy and these structural features of the network. We have also shown by real measurements that navigation in real networks is aided by hierarchical subgraphs, or scaffolds of the network.

The mechanisms of path selection and its connection with the very structure of the network is not well-understood at this moment. We argue that the in-depth understanding of the interaction between function and structure is inevitable for the deeper understanding and controlling the behavior of the networks surrounding us.

# Chapter 8

# Summary of New Results

The new results, in this work, can be divided into three groups. The first group contains the fundamentals of the function→structure approach to networks. The second and third groups unfold the steps of the function→structure analysis of networks, concerning two specific routing mechanisms, as functions.

## 1. Fundamentals of the function→structure approach to networks

**Thesis 1.1** ([49, 79])**.** *By the analysis of measurements regarding paths in networks from various areas of life (Internet, biology, air transportation, words), I have shown that, contrary to the most popular assumption, the computation of routes in networks is not due to the shortest path method.*

**Thesis 1.2** ([76, 77, 75, 78, 162])**.** *I have developed a game-theoretic model capable of handling the possible navigation schemes flexibly and unfolding their effects on the structure of the network, as follows.*

*The possible strategies of node $u \in \mathcal{P}$ is to create a set of edges: $S_u = 2^{\mathcal{P}\setminus\{u\}}$. The strategy vector $s = (s_0, s_1 \dots s_{N-1}) \in (S_0, S_1 \dots S_{N-1})$ of all the nodes defines the $G(s)$ graph as: $G(s) = \bigcup_{i=0}^{N-1}(i \times s_i)$. The cost function of node $u$ is:*

$$c_u = \sum_{\forall u \neq v} d_{G(s)}(u, v) + k(s_u), \quad u, v \in \mathcal{P} \tag{8.1}$$

*where $d_{G(s)}(u, v)$ is the communication cost from $u$ to $v$ over $G(s)$, and $k(s_u)$ is the cost of implementing the edges in $s_u$.*

## 2. The function→structure analysis of navigable networks

**Thesis 2.1** ([76])**.** *I have shown analytically that the network navigation game (NNG) contains a so-called greedy frame, which is present in all possible equilibrium states. Using analytical methods, I have given the connection probabilities between all possible node pairs in the greedy frame, which is a lower bound on the connection probability in the equilibrium state. I have given an upper bound on the connection probability analytically. Using the lower and upper bounds, I have given a general formula for the connection probability in the equilibrium state.*

**Thesis 2.2** ([76]). *I have shown analytically that the equilibrium network of the network navigation game (NNG) is sparse (average degree $< 4$), which is in good agreement with the observations in real complex networks. I have proven analytically that the equilibrium network's degree distribution is a power-law ($\gamma = 3$) in the case of the homogeneous sprinkling of the nodes. I have shown analytically that a non-homogeneous sprinkling of the nodes can adjust the exponent of the power-law. I have shown via simulations that the ($\gamma \approx 2$) case gives the lowest cost while the network is maximally navigable. I have shown with analytic approximations, that the clustering coefficient of the network is high ($\bar{c} \approx 0.45$).*

**Thesis 2.3** ([76]). *I have shown, by the analysis of real metrically embedded networks (air transportation network, human brain network, word network, Internet), that 70-90% the edges (depending on the network) of the greedy frame and the equilibrium network of the network navigation game (NNG) is also present in the real network. I have shown via simulations that the NNG can identify the critical edges, which addition or removal from the network results in the significant improvement or deterioration of navigability.*

## 3. Navigation and hierarchy

**Thesis 3.1** ([162, 161]). *I have shown analytically that the equilibrium state of the hierarchical network game (HNG) always contains a Spider graph, which is the topological consequence of the valley-free and local preference routing policies. I have shown that the model correctly predicts that peer edges appear between providers having a similar number of customers. I have validated the conclusions of the model via Internet measurements.*

**Thesis 3.2** ([49, 79, 98, 99, 65]). *I have shown by the analysis of time series recorded from human subjects, that there is an underlying hierarchy guiding human navigation in complex networked systems. I have shown that humans tend to simplify the navigation process by using a tree-like hierarchical subgraph (a scaffold) instead of the whole network.*

**Thesis 3.3** ([79, 98, 99]). *I have shown via entropy calculations that navigation based on hierarchical scaffolds can reduce the memory requirement of navigation by order of magnitude and speed up the process of learning to navigate a network from scratch compared to the shortest paths.*

# Bibliography


[1] M. Abramovitz and I. Stegun. *Handbook of Mathematical Functions*. Courier Dover Publication, 1965 (cit. on pp. 44, 48).

[2] Lada A Adamic et al. "Search in power-law networks". In: *Physical review E* 64.4 (2001), p. 046135 (cit. on pp. 21, 79).

[3] Bernhard Ager et al. "Anatomy of a large European IXP". In: *ACM SIGCOMM Computer Communication Review* 42.4 (2012), pp. 163–174 (cit. on p. 71).

[4] S. Albers et al. "On Nash equilibria for a network creation game". In: *Proc. of SODA'06*. 2006, pp. 89–98 (cit. on pp. 20, 28).

[5] Reka Albert and A.-L. Barabási. "Statistical Mechanics of Complex Networks". In: *Rev Mod Phys* 74 (2002), pp. 47–97. DOI: 10.1103/RevModPhys.74.47 (cit. on p. 13).

[6] Réka Albert, Hawoong Jeong, and Albert-László Barabási. "Error and attack tolerance of complex networks". In: *nature* 406.6794 (2000), p. 378 (cit. on p. 71).

[7] E. Anshelevich et al. "The Price of Stability for Network Design with Fair Cost Allocation". In: *Proc. of FOCS'04*. 2004, pp. 295–304 (cit. on pp. 20, 28).

[8] Andrea Avena-Koenigsberger et al. "A spectrum of routing strategies for brain networks". In: *PLoS computational biology* 15.3 (2019), e1006833 (cit. on p. 81).

[9] Daniel Awduche et al. *Overview and principles of Internet traffic engineering*. Tech. rep. 2002 (cit. on p. 71).

[10] A.-L. Barabási and Reka Albert. "Emergence of Scaling in Random Networks". In: *Science* 286 (1999), pp. 509–512. DOI: 10.1126/science.286.5439.509 (cit. on p. 13).

[11] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *science* 286.5439 (1999), pp. 509–512 (cit. on pp. 14, 18, 19).

[12] Albert-Laszlo Barabasi and Zoltan N Oltvai. "Network biology: understanding the cell's functional organization". In: *Nature reviews genetics* 5.2 (2004), p. 101 (cit. on p. 79).

[13] Albert-László Barabási and Zoltán N Oltvai. "Network biology: understanding the cell's functional organization." In: *Nat. Rev. Genet.* 5.2 (Feb. 2004), pp. 101–13. DOI: 10.1038/nrg1272 (cit. on p. 25).

[14] A. Baronchelli et al. “Networks in Cognitive Sciences”. In: *Trends in Cognitive Sciences* 17.7 (2013), pp. 348–360 (cit. on p. 64).

[15] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*. Vol. 1. Cambridge University Press Cambridge, 2008 (cit. on p. 25).

[16] Alain Barrat and Martin Weigt. “On the properties of small-world network models”. In: *The European Physical Journal B-Condensed Matter and Complex Systems* 13.3 (2000), pp. 547–560 (cit. on p. 18).

[17] Marc Barthélemy et al. “Velocity and hierarchical spread of epidemic outbreaks in scale-free networks”. In: *Phys. Rev. Lett.* 92.17 (2004), p. 178701 (cit. on p. 25).

[18] Alex Bavelas. “Communication patterns in task-oriented groups”. In: *The Journal of the Acoustical Society of America* 22.6 (1950), pp. 725–730 (cit. on p. 85).

[19] Timothy EJ Behrens et al. “What is a cognitive map? Organizing knowledge for flexible behavior”. In: *Neuron* 100.2 (2018), pp. 490–509 (cit. on pp. 25, 88).

[20] Jacob L. S. Bellmund et al. “Navigating cognition: Spatial codes for human thinking”. In: *Science* 362.6415 (2018). ISSN: 0036-8075. DOI: 10.1126/science.aat6766. eprint: http://science.sciencemag.org/content/362/6415/eaat6766.full.pdf. URL: http://science.sciencemag.org/content/362/6415/eaat6766 (cit. on pp. 25, 88).

[21] Davide Bilò et al. *Geometric Network Creation Games*. Apr. 2019 (cit. on p. 20).

[22] N. Bleistein and R.A. Handelsman. *Asymptotic Expansions of Integrals*. Dover Publications (New York), 1986 (cit. on p. 48).

[23] S Boccaletti et al. “Complex Networks: Structure and Dynamics”. In: *Phys. Rep.* 424 (2006), pp. 175–308. DOI: 10.1016/j.physrep.2005.10.009 (cit. on p. 63).

[24] M. Boguna. “Class of correlated random networks with hidden variables”. In: *Phys. Rev. E* 68 (3 2003), pp. 1–13 (cit. on p. 45).

[25] M. Boguna, F. Papadopoulos, and D. Krioukov. “Sustaining the Internet with hyperbolic mapping”. In: *Nat Comm* 1.6 (2010), pp. 1–8 (cit. on pp. 25, 67, 71).

[26] Marian Boguna, Dmitri Krioukov, and Kimberly C Claffy. “Navigability of complex networks”. In: *Nature Physics* 5.1 (2009), p. 74 (cit. on pp. 21, 25, 31, 53, 71, 79).

[27] Marián Boguñá, Fragkiskos Papadopoulos, and Dmitri Krioukov. “Sustaining the Internet with Hyperbolic Mapping”. In: *Nat. Comms.* 1 (2010), p. 62. DOI: 10.1038/ncomms1063 (cit. on p. 65).

[28] Marián Boguñá and Romualdo Pastor-Satorras. “Class of Correlated Random Networks with Hidden Variables”. In: *Phys. Rev. E* 68 (2003), p. 36112. DOI: 10.1103/PhysRevE.68.036112 (cit. on p. 45).

[29] Béla Bollobás. "Random graphs". In: *Modern graph theory*. Springer, 1998, pp. 215–252 (cit. on pp. 14, 19).

[30] E Bullmore and O Sporns. "Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems". In: *Nat. Rev. Neurosci.* 10 (2009), pp. 168–198. DOI: 10.1038/nrn2575 (cit. on p. 25).

[31] Ed Bullmore and Olaf Sporns. "Complex brain networks: graph theoretical analysis of structural and functional systems". In: *Nature Reviews Neuroscience* 10.3 (2009), p. 186 (cit. on p. 79).

[32] Guo C. et al. "BCube: a high performance, server-centric network architecture for modular data centers". In: *ACM SIGCOMM CCR* 39.4 (2009), pp. 63–74 (cit. on p. 25).

[33] CAIDA. *The CAIDA project*. http://www.caida.org (cit. on pp. 71, 77).

[34] Leila Cammoun et al. "Mapping the human connectome at multiple scales with diffusion spectrum MRI". In: *Journal of neuroscience methods* 203.2 (2012), pp. 386–397 (cit. on p. 22).

[35] José A Capitán et al. "Local-based semantic navigation on a networked representation of information". In: *PLoS ONE* 7.8 (Jan. 2012), e43694. DOI: 10.1371/journal.pone.0043694 (cit. on p. 25).

[36] Cécile Caretta Cartozo and Paolo De Los Rios. "Extended Navigability of Small World Networks: Exact Results and New Insights". In: *Phys. Rev. Lett.* 102.23 (June 2009), p. 238703. DOI: 10.1103/PhysRevLett.102.238703 (cit. on p. 25).

[37] Jean M Carlson and John Doyle. "Highly optimized tolerance: A mechanism for power laws in designed systems". In: *Physical Review E* 60.2 (1999), p. 1412 (cit. on pp. 14, 19).

[38] Claudio Castellano and Romualdo Pastor-Satorras. "Competing activation mechanisms in epidemics on networks". In: *Scientific reports* 2 (2012), p. 371 (cit. on p. 71).

[39] Miguel Castro et al. "Topology-aware routing in structured peer-to-peer overlay networks". In: *Future directions in distributed computing*. Springer, 2003, pp. 103–107 (cit. on p. 71).

[40] Dante Chialvo. "Emergent complex neural dynamics". In: *Nat. Phys.* 6.10 (Oct. 2010), pp. 744–750. DOI: 10.1038/nphys1803 (cit. on p. 25).

[41] M. Choudhury and A. Mukherjee. "The structure and dynamics of linguistic networks". In: *Dynamics on and of complex networks*. Springer, 2009, pp. 145–166 (cit. on p. 64).

[42] Elizabeth R Chrastil and William H Warren. "From cognitive maps to cognitive graphs". In: *PloS one* 9.11 (2014), e112544 (cit. on p. 88).

[43] Giulio Cimini et al. "The statistical physics of real-world networks". In: *Nature Reviews Physics* 1.1 (2019), pp. 58–71. DOI: 10.1038/s42254-018-0002-6. URL: https://doi.org/10.1038/s42254-018-0002-6 (cit. on p. 34).

[44] David Clark, William Lehr, and Steven Bauer. "Interconnection in the Internet: the policy challenge". In: (2011) (cit. on p. 71).

[45] Reuven Cohen and Shlomo Havlin. "Scale-free networks are ultrasmall". In: *Phys. Rev. Lett.* 90.5 (2003), p. 058701 (cit. on pp. 18, 53).

[46] J. Corbo and D. Parkes. "The price of selfish behavior in bilateral network formation". In: *Proc. of PODC'05*. Las Vegas, NV, USA, 2005, pp. 99–107. ISBN: 1-58113-994-2 (cit. on pp. 20, 28).

[47] Sean P Cornelius, Joo Sang Lee, and Adilson E Motter. "Dispensability of Escherichia coli's latent pathways." In: *Proc. Natl. Acad. Sci. USA* 108.8 (Feb. 2011), pp. 3124–9. DOI: 10.1073/pnas.1009772108 (cit. on p. 27).

[48] L da F Costa et al. "Characterization of complex networks: A survey of measurements". In: *Advances in physics* 56.1 (2007), pp. 167–242 (cit. on p. 14).

[49] Attila Csoma et al. "Routes obey hierarchy in complex networks". In: *Scientific reports* 7.1 (2017), pp. 1–7 (cit. on pp. 21, 79, 81, 95, 96).

[50] Raissa M D'Souza et al. "Emergence of tempered preferential attachment from optimization". In: *Proc. Natl. Acad. Sci. USA* 104.15 (2007), pp. 6112–6117 (cit. on p. 62).

[51] Alessandro Daducci et al. "The connectome mapper: an open-source processing pipeline to map connectomes with MRI". In: *PloS one* 7.12 (2012), e48121 (cit. on p. 22).

[52] E. D. Demaine et al. "The price of anarchy in network creation games". In: *Proc. of PODC '07*. 2007, pp. 292–298 (cit. on pp. 20, 28).

[53] Peter Sheridan Dodds, Roby Muhamad, and Duncan J Watts. "An experimental study of search in global social networks." In: *Science* 301.5634 (Aug. 2003), pp. 827–9. DOI: 10.1126/science.1081058 (cit. on p. 25).

[54] Peter Sheridan Dodds, Roby Muhamad, and Duncan J. Watts. "An Experimental Study of Search in Global Social Networks". In: *Science* 301.5634 (2003), pp. 827–829. ISSN: 0036-8075. DOI: 10.1126/science.1081058. eprint: http://science.sciencemag.org/content/301/5634/827.full.pdf. URL: http://science.sciencemag.org/content/301/5634/827 (cit. on p. 79).

[55] Peter Sheridan Dodds, Duncan J. Watts, and Charles F. Sabel. "Information exchange and the robustness of organizational networks". In: *Proceedings of the National Academy of Sciences* 100.21 (2003), pp. 12516–12521. ISSN: 0027-8424. DOI: 10.1073/pnas.1534702100. eprint: https://www.pnas.org/content/100/21/12516.full.pdf. URL: https://www.pnas.org/content/100/21/12516 (cit. on pp. 79, 83, 85).

[56] Christian Doerr, Norbert Blenn, and Piet Van Mieghem. "Lognormal infection times of online information spread". In: *PloS ONE* 8.5 (2013), e64349 (cit. on p. 25).

[57] S N Dorogovtsev and J F F Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford: Oxford University Press, 2003 (cit. on p. 13).

[58] Russell A Epstein et al. “The cognitive map in humans: spatial navigation and beyond”. In: *Nature neuroscience* 20.11 (2017), p. 1504 (cit. on pp. 25, 88).

[59] A. Fabrikant et al. “On a network creation game”. In: *Proc. of PODC'03*. 2003, pp. 347–351 (cit. on pp. 19, 20, 28).

[60] M Faloutsos, P Faloutsos, and C Faloutsos. “On Power-law Relationships of the {Internet} Topology”. In: *SIGCOMM*. 1999 (cit. on p. 13).

[61] D a Fell and a Wagner. “The small world of metabolism.” In: *Nat Biotechnol* 18.11 (Dec. 2000), pp. 1121–2. ISSN: 1087-0156. DOI: 10.1038/81025. URL: http://www.ncbi.nlm.nih.gov/pubmed/11062388 (cit. on p. 13).

[62] Neil Ferguson. “Capturing human behaviour”. In: *Nature* 446.7137 (2007), pp. 733–733 (cit. on p. 25).

[63] R. Ferrer i Cancho and R.V. Sole. “The small world of human language”. In: *Proc. R. Soc. Lond. B, Biological Sciences* 268.1482 (2001), pp. 2261–2265 (cit. on p. 64).

[64] R Ferrer I Cancho and R V Solé. “The small world of human language.” In: *Proc Biol Sci* 268.1482 (Nov. 2001), pp. 2261–5. ISSN: 0962-8452. DOI: 10.1098/rspb.2001.1800. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1088874%5C&tool=pmcentrez%5C&rendertype=abstract (cit. on p. 13).

[65] fit-fat-cat. *Smartphone application*. https://play.google.com/store/apps/details?id=hu.bme.tmit.lendulet.wordnavigationgame. [Online; accessed 03-20-2019]. 2016 (cit. on pp. 81, 89, 96).

[66] A. Fronczak, P. Fronczak, and J. A. Holyst. “Mean-field theory for clustering coefficients in Barabasi-Albert networks”. In: *Arxiv preprint arXiv:cond-mat/0306255 [cond-mat.stat-mech]* (2003) (cit. on p. 47).

[67] Lazaros K Gallos et al. “Scaling theory of transport in complex biological networks”. In: *Proc. Natl. Acad. Sci. USA* 104.19 (2007), pp. 7746–7751 (cit. on p. 25).

[68] Lixin Gao and Jennifer Rexford. “Stable Internet routing without global coordination”. In: *IEEE/ACM Transactions on Networking (TON)* 9.6 (2001), pp. 681–692 (cit. on pp. 74, 79).

[69] Lixin Gao and Feng Wang. “The extent of AS path inflation by routing policies”. In: *Global Telecommunications Conference, 2002. GLOBECOM'02. IEEE*. Vol. 3. IEEE. 2002, pp. 2180–2184 (cit. on p. 80).

[70] Robert S Garfinkel and George L Nemhauser. *Integer programming*. Vol. 4. Wiley New York, 1972 (cit. on p. 31).

[71] Phillipa Gill, Michael Schapira, and Sharon Goldberg. “A survey of interdomain routing policies”. In: *ACM SIGCOMM Computer Communication Review* 44.1 (2013), pp. 28–34 (cit. on p. 71).

[72] Joaquin Goni et al. "Resting-brain functional connectivity predicted by analytic measures of network communication". In: *Proc. Natl. Acad. Sci. USA* 111.2 (2014), pp. 833–8. ISSN: 1091-6490. DOI: 10.1073/pnas.1315529111 (cit. on pp. 60, 64).

[73] Temple Grandin. "Observations of cattle behavior applied to the design of cattle-handling facilities". In: *Applied Animal Ethology* 6.1 (1980), pp. 19–31 (cit. on p. 80).

[74] Albert Greenberg et al. "The cost of a cloud: research problems in data center networks". In: *ACM SIGCOMM computer communication review* 39.1 (2008), pp. 68–73 (cit. on p. 71).

[75] András Gulyás et al. "Brief announcement: network formation games can give rise to realistic networks". In: *Proceedings of the 2012 ACM symposium on Principles of distributed computing*. 2012, pp. 329–330 (cit. on p. 95).

[76] András Gulyás et al. "Navigable networks as Nash equilibria of navigation games". In: *Nature communications* 6 (2015) (cit. on pp. 95, 96).

[77] András Gulyás et al. "On Greedy Network Formation". In: *Proceedings of the SIGMETRICS W-PIN WS* (2012) (cit. on p. 95).

[78] András Gulyás et al. "On greedy network formation". In: *ACM SIGMETRICS Performance Evaluation Review* 40.2 (2012), pp. 49–52 (cit. on p. 95).

[79] András Gulyás et al. "The role of detours in individual human navigation patterns of complex networks". In: *Scientific Reports* 10.1 (2020), pp. 1–10 (cit. on pp. 95, 96).

[80] Patric Hagmann et al. "Mapping the structural core of human cerebral cortex". In: *PLoS Biol.* 6.7 (2008), pp. 1479–1493. ISSN: 15449173. DOI: 10.1371/journal.pbio.0060159 (cit. on pp. 22, 65).

[81] Denis Helic et al. "Models of human navigation in information networks based on decentralized search". In: *Proceedings of the 24th ACM conference on hypertext and social media*. ACM. 2013, pp. 89–98 (cit. on pp. 79, 83).

[82] Martijn P Van Den Heuvel et al. "Brain Communication". In: *Proc. Natl. Acad. Sci. USA* 109.28 (2012), pp. 11372–77. DOI: 10.1073/pnas.1203593109/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1203593109 (cit. on pp. 60, 64).

[83] W. Hoeffding. "Probability Inequalities for Sums of Bounded Random Variables". In: *Journal of the American Statistical Association* 58 (Mar. 1963), pp. 13–30 (cit. on p. 58).

[84] Petter Holme, Josh Karlin, and Stephanie Forrest. "An integrated model of traffic, geography and economy in the internet". In: *ACM SIGCOMM Computer Communication Review* 38.3 (2008), pp. 5–16 (cit. on p. 14).

[85] Yanqing Hu et al. "Possible Origin of Efficient Navigation in Small Worlds". In: *Phys. Rev. Lett.* 106.10 (Mar. 2011), p. 108701. DOI: 10.1103/PhysRevLett.106.108701 (cit. on p. 25).

[86] Geoff Huston. "Analyzing the Internet's BGP routing table". In: *The Internet Protocol Journal* 4.1 (2001), pp. 2–15 (cit. on p. 72).

[87] Sudarshan Iyengar et al. "A network analysis approach to understand human-wayfinding problem". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 33. 2011 (cit. on p. 79).

[88] Matthew O Jackson. "A survey of network formation models: stability and efficiency". In: *Group Formation in Economics: Networks, Clubs, and Coalitions* (2005), pp. 11–49 (cit. on p. 74).

[89] H Jeong et al. "The Large-Scale Organization of Metabolic Networks". In: *Nature* 407 (2000), pp. 651–654 (cit. on p. 13).

[90] Richard M Karp. "Reducibility among combinatorial problems". In: *Complexity of computer computations*. Springer, 1972, pp. 85–103 (cit. on p. 34).

[91] Maksim Kitsak et al. "Identification of influential spreaders in complex networks". In: *Nat. Phys.* 6.11 (2010), pp. 888–893 (cit. on p. 25).

[92] Jon Kleinberg. "Navigation in a Small World". In: *Nature* 406 (2000), p. 845. DOI: 10.1038/35022643 (cit. on pp. 25, 63).

[93] Jon M Kleinberg. "Navigation in a small world". In: *Nature* 406.6798 (2000), pp. 845–845 (cit. on pp. 21, 25, 31, 79).

[94] Jon M Kleinberg. "Small-world phenomena and the dynamics of information". In: *Advances in neural information processing systems*. 2002, pp. 431–438 (cit. on pp. 79, 83).

[95] R. Kleinberg. "Geographic routing using hyperbolic space". In: *Proc. of INFOCOM*. 2007 (cit. on p. 25).

[96] Konstantin Klemm and Victor M Eguiluz. "Growing scale-free networks with small-world behavior". In: *Physical Review E* 65.5 (2002), p. 057102 (cit. on p. 19).

[97] J. Komjáthy and K. Simon. "Generating hierarchial scale-free graphs from fractals". In: *Chaos, Solitons & Fractals* (2011) (cit. on p. 14).

[98] Attila Kőrösi et al. "A dataset on human navigation strategies in foreign networked systems". In: *Scientific data* 5 (2018), p. 180037 (cit. on pp. 21, 22, 79, 81, 89, 90, 92, 96).

[99] Attila Kőrösi et al. "fit-fat-cat dataset". In: *Open Science Framework* (2018). eprint: http://dx.doi.org/10.17605/OSF.IO/JTYVD (cit. on pp. 22, 92, 96).

[100] Dmitri Krioukov and Massimo Ostilli. "Duality between equilibrium and growing networks". In: *Phys. Rev. E* 88.2 (Aug. 2013), p. 022808. DOI: 10.1103/PhysRevE.88.022808 (cit. on p. 62).

[101] Dmitri Krioukov et al. "Hyperbolic Geometry of Complex Networks". In: *Phys. Rev. E* 82 (2010), p. 36106. DOI: 10.1103/PhysRevE.82.036106 (cit. on pp. 29, 30, 61).

[102] Dmitri Krioukov et al. "Hyperbolic geometry of complex networks". In: *Physical Review E* 82.3 (2010), p. 036106 (cit. on pp. 14, 25, 63).

[103] Simon B Laughlin and Terrence J Sejnowski. "Communication in neuronal networks". In: *Science* 301.5641 (2003), pp. 1870–1874. ISSN: 0036-8075. DOI: 10.1126/science.1089662 (cit. on pp. 60, 64).

[104] Sang Hoon Lee and Petter Holme. "A greedy-navigator approach to navigable city plans". In: *Eu. Phys. Journ. Spec. Top.* 215 (2013), pp. 135–144 (cit. on pp. 28, 62).

[105] Sang Hoon Lee and Petter Holme. "Exploring Maps with Greedy Navigators". In: *Phys. Rev. Lett.* 108.12 (Mar. 2012), p. 128701. DOI: 10.1103/PhysRevLett.108.128701 (cit. on p. 25).

[106] Sang Hoon Lee and Petter Holme. "Geometric properties of graph layouts optimized for greedy navigation". In: *Phys. Rev. E* 86.6 (Dec. 2012), p. 067103. DOI: 10.1103/PhysRevE.86.067103 (cit. on p. 25).

[107] G Li et al. "Optimal transport exponent in spatially embedded networks". In: *Phys. Rev. E* 87.4 (2013), p. 042810 (cit. on p. 63).

[108] G Li et al. "Towards design principles for optimal transport networks". In: *Phys. Rev. Lett.* 104.1 (2010), p. 018701 (cit. on p. 63).

[109] D Liben-Nowell et al. "Geographic Routing in Social Networks". In: *Proc. Natl. Acad. Sci. USA* 102 (2005), pp. 11623–11628 (cit. on p. 25).

[110] Xiang Ling et al. "Global dynamic routing for scale-free networks". In: *Physical Review E* 81.1 (2010), p. 016113 (cit. on p. 21).

[111] Aemen Lodhi, Amogh Dhamdhere, and Constantine Dovrolis. "GENESIS: An agent-based model of interdomain network formation, traffic flow and economics". In: *INFOCOM, 2012 Proceedings IEEE*. IEEE. 2012, pp. 1197–1205 (cit. on p. 14).

[112] Eng Keong Lua et al. "A survey and comparison of peer-to-peer overlay network schemes". In: *IEEE Communications Surveys & Tutorials* 7.2 (2005), pp. 72–93 (cit. on p. 71).

[113] Matthew Luckie. "Scamper: a scalable and extensible packet prober for active measurement of the internet". In: *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM. 2010, pp. 239–245 (cit. on p. 22).

[114] Zhuoqing Morley Mao et al. "Towards an Accurate AS-level Traceroute Tool". In: *Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*. SIGCOMM '03. Karlsruhe, Germany: ACM, 2003, pp. 365–378. ISBN: 1-58113-735-4. DOI: 10.1145/863955.863996. URL: http://doi.acm.org/10.1145/863955.863996 (cit. on p. 21).

[115] Sergei Maslov, Kim Sneppen, and Alexei Zaliznyak. "Detection of topological patterns in complex networks: correlation profile of the internet". In: *Physica A: Statistical Mechanics and its Applications* 333 (2004), pp. 529–540 (cit. on p. 71).

[116] Sandro Meloni, Alex Arenas, and Yamir Moreno. "Traffic-driven epidemic spreading in finite-size scale-free networks". In: *Proc. Natl. Acad. Sci. USA* 106.40 (2009), pp. 16897–16902 (cit. on p. 25).

[117] M. Mihalák and J. Schlegel. "The price of anarchy in network creation games is (mostly) constant". In: *Alg. Game Theory* (2010), pp. 276–287 (cit. on pp. 20, 28).

[118] S Milgram. "The Small World Problem". In: *Psychol. Today* 1 (1967), pp. 61–67 (cit. on p. 25).

[119] Stanley Milgram. "The small world problem". In: *Psychology today* 2.1 (1967), pp. 60–67 (cit. on p. 79).

[120] R Milo et al. "Superfamilies of Evolved and Designed Networks". In: *Science* 303 (2004), pp. 1538–1542. DOI: 10.1126/science.1089167 (cit. on p. 65).

[121] Giovanna Miritello, Esteban Moro, and Rubén Lara. "Dynamical strength of social ties in information spreading". In: *Phys. Rev. E* 83.4 (2011), p. 045102 (cit. on p. 25).

[122] Wenke Möhring, Andrea Frick, and Nora S. Newcombe. "Spatial scaling, proportional thinking, and numerical understanding in 5- to 7-year-old children". In: *Cognitive Development* 45 (2018), pp. 57–67. ISSN: 0885-2014. DOI: https://doi.org/10.1016/j.cogdev.2017.12.001. URL: http://www.sciencedirect.com/science/article/pii/S0885201417300011 (cit. on pp. 25, 88).

[123] J M Montoya and R V Solé. *Small World Patterns in Food Webs*. Technical Report 00-10-059. Santa Fe Institute, 2000 (cit. on p. 13).

[124] José M Montoya, Stuart L Pimm, and Ricard V Solé. "Ecological networks and their fragility". In: *Nature* 442.7100 (2006), p. 259 (cit. on p. 79).

[125] Yamir Moreno, Maziar Nekovee, and Amalio F Pacheco. "Dynamics of rumor spreading in complex networks". In: *Phys. Rev. E* 69.6 (2004), p. 066130 (cit. on p. 25).

[126] Lev Muchnik et al. "Origins of power-law degree distribution in the heterogeneity of human activity in social networks." In: *Sci. Rep.* 3 (2013), p. 1783. ISSN: 2045-2322. DOI: 10.1038/srep01783 (cit. on p. 62).

[127] Kevin Murphy et al. "The impact of global signal regression on resting state correlations: are anti-correlated networks introduced?" In: *Neuroimage* 44.3 (2009), pp. 893–905 (cit. on p. 23).

[128] Nora Newcombe. "Harnessing Spatial Thinking to Support Stem Learning". In: *OECD Education Working Papers* 161 (2017). DOI: https://doi.org/https://doi.org/10.1787/7d5dcae6-en. URL: https://www.oecd-ilibrary.org/content/paper/7d5dcae6-en (cit. on pp. 25, 88).

[129] Nora S. Newcombe. "Individual variation in human navigation". In: *Current Biology* 28.17 (2018), R1004–R1008. ISSN: 0960-9822. DOI: https://doi.org/10.1016/j.cub.2018.04.053. URL: http://www.sciencedirect.com/science/article/pii/S0960982218305256 (cit. on p. 88).

[130] M E J Newman. "Power Laws, Pareto Distributions and Zipf's Law". In: *Contemp. Phys.* 46.5 (2005), pp. 323–351. DOI: 10.1080/00107510500052444 (cit. on p. 45).

[131] M E J Newman. "The Structure and Function of Complex Networks". In: *SIAM Rev.* 45.2 (2003), pp. 167–256. DOI: 10.1137/S003614450342480 (cit. on p. 63).

[132] M E J Newman. "The Structure of Scientific Collaboration Networks". In: *Proc. Natl. Acad. Sci. USA* 98 (2001), pp. 404–409 (cit. on p. 13).

[133] Mark Newman. *Networks: an introduction*. Oxford university press, 2010 (cit. on p. 21).

[134] Mark EJ Newman. "The structure and function of complex networks". In: *SIAM review* 45.2 (2003), pp. 167–256 (cit. on pp. 8, 13).

[135] N. Nisan. *Algorithmic game theory*. Cambridge Universiy Press, 2007 (cit. on pp. 20, 28).

[136] Jae Dong Noh and Heiko Rieger. "Random walks on complex networks". In: *Physical review letters* 92.11 (2004), p. 118701 (cit. on p. 21).

[137] OpenFlights. *Airport Database*. `www.openflights.org`. [Online; accessed 05-06-2016] (cit. on p. 22).

[138] Christos H. Papadimitriou and David Ratajczak. "On a conjecture related to geometric routing". In: *Theor. Comput. Sci.* 344.1 (Nov. 2005), pp. 3–14. DOI: 10.1016/j.tcs.2005.06.022 (cit. on p. 28).

[139] F. Papadopoulos et al. "Greedy Forwarding in Dynamic Scale-Free Networks Embedded in Hyperbolic Metric Spaces". In: *Proc. of IEEE Infocom*. IEEE. 2010, pp. 1–9 (cit. on pp. 45, 52).

[140] Fragkiskos Papadopoulos, Constantinos Psomas, and Dmitri Krioukov. "Network mapping by replaying hyperbolic growth". In: *IEEE ACM T Netw* (2014). DOI: 10.1109/TNET.2013.2294052 (cit. on p. 65).

[141] Fragkiskos Papadopoulos et al. "Popularity versus similarity in growing networks". In: *Nature* 489 (Sept. 2012), pp. 537–540. DOI: 10.1038/nature11459 (cit. on pp. 29, 61, 64, 65).

[142] Romualdo Pastor-Satorras and Alessandro Vespignani. "Epidemic spreading in scale-free networks". In: *Phys. Rev. Lett.* 86.14 (2001), p. 3200 (cit. on p. 25).

[143] Mukaddim Pathan and Rajkumar Buyya. "A taxonomy of CDNs". In: *Content delivery networks*. Springer, 2008, pp. 33–77 (cit. on p. 71).

[144] M Penrose. *Random Geometric Graphs*. Oxford: Oxford University Press, 2003 (cit. on p. 29).

[145] Jonathan D Power et al. "Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion". In: *Neuroimage* 59.3 (2012), pp. 2142–2154 (cit. on p. 23).

[146] Yakov Rekhter, Tony Li, and Susan Hares. *A border gateway protocol 4 (BGP-4)*. Tech. rep. 2005 (cit. on p. 71).

[147] Chris J Rhodes and Roy M Anderson. "Power laws governing epidemics in isolated populations". In: *Nature* 381.6583 (1996), pp. 600–602 (cit. on p. 25).

[148] Rome2Rio. *Flights Database*. `https://www.rome2rio.com/`. [Online; accessed 05-07-2016] (cit. on p. 22).

[149] Vittorio Rosato et al. "Is the topology of the Internet network really fit to sustain its function?" In: *Physica A: Statistical Mechanics and its Applications* 387.7 (2008), pp. 1689–1704 (cit. on p. 71).

[150] Martin Rosvall et al. "Memory in network flows and its effects on spreading dynamics and community detection". In: *Nature communications* 5 (2014), p. 4630 (cit. on pp. 81, 89).

[151] Hernán D Rozenfeld, Chaoming Song, and Hernán A Makse. "Small-world to fractal transition in complex networks: a renormalization group approach". In: *Phys. Rev. Lett.* 104.2 (2010), p. 025701 (cit. on p. 63).

[152] Ratnasamy S. et al. "A scalable content-addressable network". In: *Proc. of SIGCOMM '01*. San Diego, California, United States, 2001, pp. 161–172 (cit. on p. 25).

[153] Vsevolod Salnikov, Michael T Schaub, and Renaud Lambiotte. "Using higher-order Markov models to reveal flow-based communities in networks". In: *Scientific reports* 6 (2016), p. 23194 (cit. on p. 89).

[154] Claude Elwood Shannon. "A mathematical theory of communication". In: *Bell system technical journal* 27.3 (1948), pp. 379–423 (cit. on p. 85).

[155] Yuval Shavitt and Eran Shir. "DIMES: Let the Internet measure itself". In: *ACM SIGCOMM Computer Communication Review* 35.5 (2005), pp. 71–74 (cit. on p. 71).

[156] Ozgür Simsek and David Jensen. "Navigating networks by using homophily and degree." In: *Proc. Natl. Acad. Sci. USA* 105.35 (Sept. 2008), pp. 12758–62. DOI: 10.1073/pnas.0800497105 (cit. on pp. 21, 25).

[157] Özgür Şimşek and David Jensen. "Navigating networks by using homophily and degree". In: *Proceedings of the National Academy of Sciences* 105.35 (2008), pp. 12758–12762 (cit. on p. 79).

[158] Philipp Singer et al. "Detecting memory and structure in human navigation patterns using markov chain models of varying order". In: *PloS one* 9.7 (2014), e102070 (cit. on p. 89).

[159] Ion Stoica et al. "Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications". In: *Proc. of SIGCOMM'01*. San Diego, California, Aug. 2001 (cit. on p. 25).

[160] I. Stojmenovic. "Position-based routing in ad hoc networks". In: *IEEE Communications Magazine* (2002) (cit. on p. 25).

[161] Dávid Szabó and András Gulyás. "Notes on the Topological Consequences of BGP Policy Routing on the Internet AS Topology". In: *Advances in Communication Networking*. Springer Berlin Heidelberg, 2013, pp. 274–281 (cit. on pp. 62, 96).

[162] David Szabo et al. "Deductive way of reasoning about the internet AS level topology". In: *Chinese Physics B* 24.11 (2015), p. 118901 (cit. on pp. 95, 96).

[163] *The Cooperative Association for Internet Data Analysis (CAIDA)* (cit. on p. 22).

[164] P. H E Tiesinga et al. "Optimal information transfer in synchronized neocortical neurons". In: *Neurocomputing* 38-40 (2001), pp. 397–402. ISSN: 09252312. DOI: 10.1016/S0925-2312(01)00464-7 (cit. on pp. 60, 64).

[165] J. Travers and S. Milgram. "An Experimental Study of the Small World Problem". In: *Sociometry* 32 (1969), pp. 425–443 (cit. on p. 24).

[166] J Travers and S Milgram. "An Experimental Study of the Small World Problem". In: *Sociometry* 32 (1969), pp. 425–443 (cit. on p. 25).

[167] *University of Oregon RouteViews Project* (cit. on pp. 22, 71).

[168] John Glen Wardrop. "Some theoretical aspects of road traffic research". In: *Inst Civil Engineers Proc London/UK/.* 1952 (cit. on p. 79).

[169] Duncan J Watts, Peter Sheridan Dodds, and Mark EJ Newman. "Identity and search in social networks". In: *Science* 296.5571 (2002), pp. 1302–1305 (cit. on pp. 21, 25, 31, 79, 82).

[170] Duncan J Watts and Steven H Strogatz. "Collective dynamics of 'small-world'networks". In: *Nature* 393.6684 (1998), p. 440 (cit. on pp. 13, 14, 16, 79, 81).

[171] Robert West and Jure Leskovec. "Human wayfinding in information networks". In: *Proceedings of the 21st international conference on World Wide Web*. ACM. 2012, pp. 619–628 (cit. on p. 79).

[172] WordFind. *Three Letter English Word Database*. http://www.wordfind.com/3-letter-words/. [Online; accessed 10-10-2016] (cit. on p. 22).

[173] Bo Xiao et al. "Modeling the IPv6 internet AS-level topology". In: *Physica A: Statistical Mechanics and its Applications* 388.4 (2009), pp. 529–540 (cit. on p. 71).

[174] Takuji Yamada and Peer Bork. "Evolution of biomolecular networks: lessons from metabolic and protein interactions." In: *Nat. Rev. Mol. Cell. Bio.* 10.11 (Nov. 2009), pp. 791–803. DOI: 10.1038/nrm2787 (cit. on p. 25).

[175] Keyou You, Roberto Tempo, and Li Qiu. "Distributed algorithms for computation of centrality measures in complex networks". In: *IEEE Transactions on Automatic Control* 62.5 (2017), pp. 2080–2094 (cit. on p. 87).

[176] Wayne W Zachary. "An information flow model for conflict and fission in small groups". In: *Journal of anthropological research* 33.4 (1977), pp. 452–473 (cit. on p. 13).

[177] Shanjiang Zhu and David Levinson. "Do people use the shortest path? An empirical test of Wardrop's first principle". In: *PloS one* 10.8 (2015) (cit. on p. 80).

# List of Symbols

| | |
|---|---|
| $\bar{k}$ | Average degree of the network |
| $\gamma$ | Exponent of the power-law degree distribution |
| $C$ | Global clustering coefficient of a network |
| $c_i$ | Local clustering coefficient of node $i$ |
| $D$ | Diameter of a network |
| $k_i$ | Degree of node $k$ |
| $N$ | Number of nodes in a network |
| $P(k)$ | Degree distribution of a network |